OPIC
OFFICE DE LA PROPRIÉTÉ INTELLECTUELLE DU CANADA

CIPO
CANADIAN INTELLECTUAL PROPERTY OFFICE

(12) (19) (CA) **Demande-Application**

(21) (A1) **2,211,018**
(22) 1997/09/19
(43) 1998/03/30

(72) GIJZEN, Mark, CA
(71) Her Majesty in Right of Canada as represented by the Minister of Agriculture and Agri-Food Canada, CA

(51) Int.Cl.[6] C12N 15/53, A01H 5/00, C12N 15/82
(30) 1996/09/30 (2,186,833) CA
(30) 1996/09/30 (08/723,414) US

(54) **REGION DE REGULATION DE L'ADN POUR LA PEROXYDASE DES TEGUMENTS**

(54) **SEED COAT DNA REGULATORY REGION AND PEROXIDASE**

(57) Caractérisation et présentation d'une nouvelle séquence génomique spécifique pour le tégument. Les régions régulatrices voisines de l'ADN ont également été caractérisées. Le peroxydase de tégument est traduit sous forme de protéine précurseur de 38 kDa, à 352 acides aminés, renfermant une séquence-signal de 26 acides aminés; elle donne, par clivage, une protéine de 35 kDa. Les plantes renfermant un allèle Ep dominant accumulent de grandes quantités de peroxydase dans les cellules sabliers du subépiderme. Les génotypes epep homozygotes récessifs n'accumulent pas de peroxydase dans ces cellules et leur part dans l'activité totale de la peroxydase du tégument se trouve sensiblement réduite. Les sondes dérivées de l'ADNc ou de l'ADN génomique peuvent servir à déceler les polymorphismes qui distinguent les génotypes EpEp et epep. La coségrégation des polymorphismes dans une population F_2 provenant d'un croisement de plantes EpEp et epep montre que le locus Ep code la protéine peroxydase. Une comparaison des allèles Ep et ep révèle qu'il manque 87 bp dans le gène récessif pour le codon initial de traduction. L'expression hétérologue ainsi que les vecteurs et les hôtes utilisés pour l'expression de la peroxydase du tégument sont également présentés. La région régulatrice de l'ADN spécifique pour la semence peut servir à contrôler l'expression i) de certains gènes, comme ceux codant la résistance aux herbicides, ii) de protéines virales du tégument, protégeant contre l'infection, iii) de protéines à intérêt commercial (p. ex. en pharmacie), iv) de protéines modifiant la valeur nutritive, le goût ou le conditionnement des semences; enfin, elle peut servir à v) éliminer biologiquement des insectes ou des agents pathogènes (p. ex. B. thuringiensis).

(57) A novel seed coat specific peroxidase genomic sequence is characterized and presented. Adjacent DNA regulatory regions have also been characterized. The seed coat peroxidase is translated as a 352 amino acid precursor protein of 38 kDa comprising a 26 amino acid signal sequence which when cleaved results in a 35 kDa protein. Plants containing a dominant Ep allele accumulate large amounts of peroxidase in the hourglass cells of the subepidermis. Homozygous recessive epep genotypes do not accumulate peroxidase in the hourglass cells and are much reduced in total seed coat peroxidase activity. Probes derived from the cDNA, or genomic DNA can be used to detect polymorphisms that distinguished EpEp and epep genotypes. Cosegregation of the polymorphisms in an F_2 population from a cross of EpEp and epep plants shows that the Ep locus encodes the seed coat peroxidase protein. Comparison of Ep and ep alleles indicates that the recessive gene lacks 87 bp of sequence encompassing the translation start codon. The heterologous expression, as well as vectors and hosts to be used for the expression of the seed coat peroxidase, are also disclosed. The seed-specific DNA regulatory region may be used to control expression of genes of interest such as i) genes encoding herbicide resistance, or ii) biological control of insects or pathogens (e.g. B. thuringiensis), or iii) viral coat proteins to protect against viral infections, or iv) proteins of commercial interest (e.g. pharmaceutical), and v) proteins that alter the nutritive value, taste, or processing of seeds.

ABSTRACT OF THE DISCLOSURE

A novel seed coat specific peroxidase genomic sequence is characterized and presented. Adjacent DNA regulatory regions have also been characterized. The seed coat peroxidase is translated as a 352 amino acid precursor protein of 38 kDa comprising a 26 amino acid signal sequence which when cleaved results in a 35 kDa protein. Plants containing a dominant *Ep* allele accumulate large amounts of peroxidase in the hourglass cells of the subepidermis. Homozygous recessive *epep* genotypes do not accumulate peroxidase in the hourglass cells and are much reduced in total seed coat peroxidase activity. Probes derived from the cDNA, or genomic DNA can be used to detect polymorphisms that distinguished *EpEp* and *epep* genotypes. Cosegregation of the polymorphisms in an F_2 population from a cross of *EpEp* and *epep* plants shows that the *Ep* locus encodes the seed coat peroxidase protein. Comparison of *Ep* and *ep* alleles indicates that the recessive gene lacks 87 bp of sequence encompassing the translation start codon. The heterologous expression, as well as vectors and hosts to be used for the expression of the seed coat peroxidase, are also disclosed. The seed-specific DNA regulatory region may be used to control expression of genes of interest such as i) genes encoding herbicide resistance, or ii) biological control of insects or pathogens (e.g. B. thuringiensis), or iii) viral coat proteins to protect against viral infections, or iv) proteins of commercial interest (e.g. pharmaceutical), and v) proteins that alter the nutritive value, taste, or processing of seeds.

SEED COAT SPECIFIC DNA REGULATORY REGION AND PEROXIDASE

The present invention relates to a novel DNA molecule comprising a plant seed coat specific DNA regulatory region and a novel structural gene encoding a peroxidase. The seed-coat specific DNA regulatory region may also be used to control the expression of other genes of interest within the seed coat.

BACKGROUND OF THE INVENTION

Full citations for references appear at the end of the Examples section.

Peroxidases are enzymes catalyzing oxidative reactions that use H_2O_2 as an electron acceptor. These enzymes are widespread and occur ubiquitously in plants as isozymes that may be distinguished by their isoelectric points. Plant peroxidases contribute to the structural integrity of cell walls by functioning in lignin biosynthesis and suberization, and by forming covalent cross-linkages between extension, cellulose, pectin and other cell wall constituents (Campa, 1991). Peroxidases are also associated with plant defence responses and resistance to pathogens (Bowles, 1990; Moerschbacher 1992). Soybeans contain 3 anionic isozymes of peroxidase with a minimum M_r of 37 kDa (Sessa and Anderson, 1981). Recently one peroxidase isozyme, localised within the seed coat of soybean, has been characterized with a M_r of 37 kDa (Gillikin and Graham, 1991).

In an analysis of soybean seeds, Buttery and Buzzell (1968) showed that the amount of peroxidase activity present in seed coats may vary substantially among different cultivars. The presence of a single dominant gene *Ep* causes a high seed coat peroxidase phenotype (Buzzell and Buttery, 1969). Homozygous recessive *epep* plants are ~100-fold lower in seed coat peroxidase activity. This results from a reduction in the amount of peroxidase enzyme present, primarily in the hourglass cells of the subepidermis (Gijzen *et al.*, 1993). In plants carrying the *Ep* gene, peroxidase is heavily concentrated in the hourglass cells (osteosclereids). These cells form a highly differentiated cell layer with thick, elongated secondary walls and large intercellular spaces (Baker *et al.*, 1987). Hourglass cells develop between the epidermal macrosclereids and the underlying articulated parenchyma, and are a prominent feature of seed coat anatomy at full maturity. The cytoplasm exudes from the hourglass cells upon imbibition with water and a distinct peroxidase isozyme constitutes five to 10% of the total soluble protein in *EpEp* seed coats. It is not known why the hourglass cells accumulate large amounts of peroxidase, but the sheer abundance and relative purity of the enzyme in soybean seed coats is significant because peroxidases are versatile enzymes with many commercial and industrial applications. Studies of soybean seed coat peroxidase have shown this enzyme to have useful catalytic properties and a high degree of thermal stability even at extremes of pH (McEldoon *et al.*, 1995). These properties result in the preferred use of soybean peroxidase, over that of horseradish peroxidase, in diagnostic assays as an enzyme label for antigens, antibodies, oligonucleotide probes, and within staining techniques. Johnson et al report on the use of soybean peroxidase for the deinking of printed waste paper (U.S. 5,270,770;

December 6, 1994) and for the biocatalytic oxidation of primary alcohols (U.S. 5,391,488; February 13, 1996). Soybean peroxidase has also been used as a replacement for chlorine in the pulp and paper industry, or as formaldehyde replacement (Freiberg, 1995).

An anionic soybean peroxidase from seed coats has been purified (Gillikin and Graham, 1991). This protein has a pI of 4.1 and M_r of 37 kDa. A method for the bulk extraction of peroxidase from seed hulls of soybean using a freeze thaw technique has also been reported (U.S. 5,491,085, February 13, 1996, Pokara and Johnson).

Lagrimini et al (1987) disclose the cloning of a ubiquitous anionic peroxidase in tobacco encoding a protein of M_r of 36 kDa. This peroxidase has also been over expressed in transgenic tobacco plants (Lagrimini et al 1990) and Maliyakal discloses the expression of this gene in cotton (WO 95/08914).

Huangpu et al (1995) reported the partial cloning of a soybean anionic seed coat peroxidase. The 1031 bp sequence contained an open reading frame of 849 bp encoding a 283 amino acid protein with a Mr of 30,577. The M_r of this peroxidase is 7 kDa less than what one would expect for a soybean seed coat peroxidase as reported by Gillikin and Graham (1991) and possibly represents another peroxidase isozyme within the seed coat.

The upstream promoter sequences for two poplar peroxidases have been described by Osakabe et al (1995). A number of characteristic regulatory sites were identified from comparison of these sequences to existing promoter elements. Additionally, a cryptic promoter with apparent specificity for seed coat tissues was isolated from tobacco by a promoter trapping strategy (Fobert et al. 1994). The upstream regulatory sequences associated with the Ep gene in soybean are distinct from these and other previously characterized promoters. The soybean Ep promoter drives high-level expression in a cell and tissue specific manner. The peroxidase protein encoded by the Ep gene accumulates in the seed coat tissues, especially in the hour glass cells of the subepidermis. Minimal expression of the gene is detected in root tissues.

One problem arising from the desired use of soybean seed coat peroxidase is that there is variability between soybean varieties regarding peroxidase production (Buttery and Buzzell, 1986; Freiberg, 1995). Due to the commercial interest in the use of soybean seed coat peroxidase new methods of producing this enzyme are required. Therefore, the gene responsible for the expression of the 37 kDa isozyme in soybean seed coat was isolated and characterized.

Furthermore, novel regulatory regions obtained from the genomic DNA of soybean seed coat peroxidase have been isolated and characterized and are useful in directing the expression of genes of interest in seed coat tissues.

SUMMARY OF THE INVENTION

The present invention relates to a DNA molecule that encodes a soybean seed coat peroxidase and associated DNA regulatory regions.

This invention also embraces isolated DNA molecules comprising the nucleotide sequence of either SEQ ID NO:1 (the cDNA encoding soybean seed coat peroxidase) SEQ ID No:2 (the genomic sequence).

This invention also provides for a chimeric DNA molecule comprising a seed coat-specific regulatory region having nucleotides 1-1532 of SEQ ID NO:2 and a gene of interest under control of this DNA regulatory region. Also included within this invention are chimeric DNA molecules comprising genomic DNA sequences exemplified by nucleotides 412-1041, 1234-2263 or 2430-2691 of SEQ ID NO:2. Furthermore, this invention is directed to isolated DNA molecules comprising at least

1) 24 contiguous nucleotides selected from nucleotides 1-1532 of SEQ ID NO:2;
2) 32 contiguous nucleotides selected from nucleotides 412-1041 of SEQ ID NO:2;
3) 23 contiguous nucleotides selected from nucleotides 1234-2263 of SEQ ID NO:2; or
4) 22 contiguous nucleotides selected from nucleotides 2430-2691 of SEQ ID NO:2.

The present invention also provides for vectors which comprise DNA molecules encoding soybean seed coat peroxidase. Such a construct may include the DNA regulatory region from SEQ ID NO:2, including nucleotides 1-1532, or at least 24 contiguous nucleotides selected from nucleotides 1-1532 of SEQ ID NO:2 in conjunction with the seed coat peroxidase gene, or the seed coat peroxidase gene under the control of any suitable constitutive or inducible promoter of interest.

This invention is also directed towards vectors which comprise a gene of interest placed under the control of a DNA regulatory element derived from the genomic sequence encoding soybean seed coat peroxidase. Such a regulatory element includes nucleotides 1-1532 of SEQ ID NO:2, or at least 24 contiguous nucleotides selected from nucleotides 1-1532 of SEQ ID NO:2. Elements comprising nucleotides 412-1041, 1234-2263 or 2430-2691 of SEQ ID NO:2, or 32 contiguous nucleotides selected from nucleotides 412-1041 of SEQ ID NO:2, 23 contiguous nucleotides selected from nucleotides 1234-2263 of SEQ ID NO:2, or 22 contiguous nucleotides selected from nucleotides 2430-2691 of SEQ ID NO:2 may also be used.

This invention also embraces prokaryotic and eukaryotic cells comprising the vectors identified above. Such cells may include bacterial, insect, mammalian, and plant cell cultures.

This invention also provides for transgenic plants comprising the seed coat peroxidase gene under control of constitutive or inducible promoters. Furthermore,

this invention also relates to transgenic plants comprising the DNA regulatory regions of nucleotides 1-1532 of SEQ ID NO:2 controlling a gene of interest, or comprising genes of interest in functional association with genomic DNA sequences exemplified by nucleotides 412-1041, 1234-2263 or 2430-2691 of SEQ ID NO:2. Also embraced by this invention are transgenic plants having regulatory regions comprising at least 24 contiguous nucleotides selected from nucleotides 1-1532 of SEQ ID NO:2, 32 contiguous nucleotides selected from nucleotides 412-1041 of SEQ ID NO:2, 23 contiguous nucleotides selected from nucleotides 1234-2263 of SEQ ID NO:2, or 22 contiguous nucleotides selected from nucleotides 2430-2691 of SEQ ID NO:2.

This invention is also directed to a method for the production of soybean seed coat peroxidase in a host cell comprising:

i) transforming the host cell with a vector comprising an oligonucleotide sequence that encodes soybean seed coat peroxidase; and

ii) culturing the host cell under conditions to allow expression of the soybean seed coat peroxidase.

This invention also provides for a process for producing a heterologous gene of interest within seed coats of a transformed plant, comprising propagating a plant transformed with a vector comprising a gene of interest under the control of nucleotides 1-1532 of SEQ ID NO:2. Furthermore, this invention embraces a process for producing a heterologous gene of interest within seed coats of a transformed plant, comprising propagating a plant transformed with a vector comprising a gene of interest

under the control of a regulatory region comprising at least 24 nucleotides selected from nucleotides 1-1532 of SEQ ID NO:2.

Although the present invention is exemplified by a soybean seed coat peroxidase and adjacent DNA regulatory regions, in practice any gene of interest can be placed downstream from the DNA regulatory region for seed coat specific expression.

BRIEF DESCRIPTION OF THE DRAWINGS

These and other features of the invention will become more apparent from the following description in which reference is made to the appended drawings wherein:

Figure 1 is the cDNA and deduced amino acid sequence of soybean seed coat peroxidase. Nucleotides are numbered by assigning +1 to the first base of the ATG start codon; amino acids are numbered by assigning +1 to the N-terminal Gln residue after cleavage of the putative signal sequence. The N-terminal signal sequence, the region of the active site, and the heme-binding domain are underlined. The numerals I, II and III placed directly above single nucleotide gaps in the sequence indicate the three intron splice positions. The target site and direction of five different PCR primers are shown with dotted lines above the nucleotide sequence. An asterisk (*) marks the translation stop codon.

Figure 2 is the genomic DNA sequence of the Soybean seed coat peroxidase.

Figure 3 is a comparison of soybean seed coat peroxidase with other closely related plant peroxidases. The GenBank accession numbers are provided next to the name of the plant from which the peroxidase was isolated. The accession number for the soybean sequence is L78163. (A) A comparison of the nucleic acid sequences; (B) A comparison of the amino acid sequences.

Figure 4 is a restriction fragment length polymorphisms between *EpEp* and *epep* genotypes using the seed coat peroxidase cDNA as probe. Genomic DNA of soybean lines OX312 (*epep*) and OX347 (*EpEp*) was digested with restriction enzyme, separated by electrophoresis in a 0.5% agarose gel, transferred to nylon, and hybridized with ^{32}P-labelled cDNA encoding the seed coat peroxidase. The size of the hybridizing fragments was estimated by comparison to standards and is indicated on the right.

Figure 5 exhibits the structure of the *Ep* Locus. A 17 kb fragment including the *Ep* locus is illustrated schematically. A 3.3 kb portion of the gene is enlarged and exons and introns are represented by shaded and open boxes, respectively. The final enlargement of the 5' region shows the location and DNA sequence around the 87 bp deletion occurring in the *ep* allele of soybean line OX312. Nucleotides are numbered by assigning +1 to the first base of the ATG start codon.

Figure 6 displays PCR analysis of *EpEp* and *epep* genotypes using primers derived from the seed coat peroxidase cDNA. Genomic DNA from soybean lines OX312 (*epep*) and OX347 (*EpEp*) was used as template for PCR analysis with four different primer sets. Amplification products were separated by electrophoresis through a 0.8% agarose gel and visualized under UV light after staining with ethidium bromide. Genotype and primer combinations are

indicated at the top of the figure. The size in base pairs of the amplified DNA fragments are indicated on the right.

Figure 7 exhibits PCR analysis of an F2 population from a cross of *EpEp* and *epep* genotypes. Genomic DNA was used as template for PCR analysis of the parents (P) and 30 F_2 individuals. The cross was derived from the soybean lines OX312 (*epep*) and OX347 (*EpEp*). Plants were self pollinated and seeds were collected and scored for seed coat peroxidase activity. The symbols (-) and (+) indicate low and high seed coat peroxidase activity, respectively. Primers prx9+ and prx10- were used in the amplification reactions. Products were separated by electrophoresis through a 0.8% agarose gel and visualized under UV light after staining with ethidium bromide. The migration of molecular markers and their corresponding size in kb is also shown (lanes M).

Figure 8 displays PCR analysis of six different soybean cultivars with primers derived from the seed coat peroxidase cDNA sequence. Genomic DNA was used as template for PCR analysis of three *EpEp* cultivars and three *epep* cultivars. Primers used in the amplification reactions and the size of the DNA product is indicated on the left. Products were separated by electrophoresis through a 0.8% agarose gel and visualized under UV light after staining with ethidium bromide.

(A) Forward and reverse primers are downstream from deletion

(B) Forward primer anneals to site within deletion

(C) Primers span deletion

Figure 9 shows the accumulation of peroxidase RNA in tissues of GEp and *epep* plants. **Figure 9(A)**: A comparison of peroxidase transcript abundance in cultivars Harosoy 63 (Ep) or Marathon (ep). Seed and pod tissues were sampled at a late stage of development corresponding to a whole seed fresh weight of 250 mg. Root and leaf tissue was from six week old plants. Autoradiograph exposed for 96 h. **Figure 9(B)**: Developmental expression of peroxidase in cultivar Harosoy 63 (Ep). Flowers were sampled immediately after opening. Seed coat tissues were sampled at four stages of development corresponding to a whole seed fresh weight of: lane 1, 50 mg; lane 2, 100 mg; lane 3, 200 mg; lane 4, 250 mg. Autoradiograph exposed for 20 h.

DESCRIPTION OF PREFERRED EMBODIMENT

The present invention is directed to a novel oligonucleotide sequence encoding a seed coat peroxidase and associated DNA regulatory regions.

According to the present invention DNA sequences that are "substantially homologous" includes sequences that are identified under conditions of high stringency. "High stringency" refers to Southern hybridization conditions employing washes at 65°C with 0.1 x SSC, 0.5 % SDS.

By "DNA regulatory region" it is meant any region within a genomic sequence that has the property of controlling the expression of a DNA sequence that is operably linked with the regulatory region. Such regulatory regions may include promoter or enhancer regions, and other regulatory elements recognized by one of skill in the art. A segment of the DNA regulatory region is exemplified in this invention, however, as is understood by one of skill in the art, this region may be used as a probe to identify surrounding regions involved in the regulation of adjacent DNA, and such surrounding regions are also included within the scope of this invention.

In the context of this disclosure, the term "promoter" or "promoter region" refers to a sequence of DNA, usually upstream (5') to the coding sequence of a structural gene, which controls the expression of the coding region by providing the recognition for RNA polymerase and/or other factors required for transcription to start at the correct site.

There are generally two types of promoters, inducible and constitutive. An "inducible promoter" is a promoter that is capable of directly or indirectly activating transcription of one or more DNA sequences or genes in response to an inducer. In the absence of an inducer the DNA sequences or genes will not be transcribed. Typically the protein factor, that binds specifically to an inducible promoter to activate transcription, is present in an inactive form which is then directly or indirectly converted to the active form by the inducer. The inducer can be a chemical agent such as a protein, metabolite, growth regulator, herbicide or phenolic compound or a physiological stress imposed directly by heat, cold, salt, or toxic elements or indirectly through the action of a pathogen or disease agent such as a virus. A plant cell containing an inducible promoter may be exposed to an inducer by externally applying the inducer to the cell or plant such as by spraying, watering, heating or similar methods.

By "constitutive promoter" it is meant a promoter that directs the expression of a gene throughout the various parts of a plant and continuously throughout plant development. Examples of known constitutive promoters include those associated with the CaMV 35S transcript and *Agrobacterium* Ti plasmid nopaline synthase gene.

The chimeric gene constructs of the present invention can further comprise a 3' untranslated region. A 3' untranslated region refers to that portion of a gene comprising a DNA segment that contains a polyadenylation signal and any other regulatory signals capable of effecting mRNA processing or gene expression. The

polyadenylation signal is usually characterized by effecting the addition of polyadenylic acid tracks to the 3' end of the mRNA precursor. Polyadenylation signals are commonly recognized by the presence of homology to the canonical form 5' AATAAA-3' although variations are not uncommon.

Examples of suitable 3' regions are the 3' transcribed non-translated regions containing a polyadenylation signal of *Agrobacterium* tumour inducing (Ti) plasmid genes, such as the nopaline synthase (*Nos* gene) and plant genes such as the soybean storage protein genes and the small subunit of the ribulose-1, 5-bisphosphate carboxylase (ssRUBISCO) gene. The 3' untranslated region from the structural gene of the present construct can therefore be used to construct chimeric genes for expression in plants.

The chimeric gene construct of the present invention can also include further enhancers, either translation or transcription enhancers, as may be required. These enhancer regions are well known to persons skilled in the art, and can include the ATG initiation codon and adjacent sequences. The initiation codon must be in phase with the reading frame of the coding sequence to ensure translation of the entire sequence. The translation control signals and initiation codons can be from a variety of origins, both natural and synthetic. Translational initiation regions may be provided from the source of the transcriptional initiation region, or from the structural gene. The sequence can also be derived from the promoter selected to express the gene, and can be specifically modified so as to increase translation of the mRNA.

To aid in identification of transformed plant cells, the constructs of this invention may be further manipulated to include plant selectable markers. Useful selectable markers include enzymes which provide for resistance to an antibiotic such as gentamycin, hygromycin, kanamycin, and the like. Similarly, enzymes providing for production of a compound identifiable by colour change such as *GUS* (β-glucuronidase), or luminescence, such as luciferase are useful.

Also considered part of this invention are transgenic plants containing the chimeric gene construct of the present invention. Methods of regenerating whole plants from plant cells are known in the art, and the method of obtaining transformed and regenerated plants is not critical to this invention. In general, transformed plant cells are cultured in an appropriate medium, which may contain selective agents such as antibiotics, where selectable markers are used to facilitate identification of transformed plant cells. Once callus forms, shoot formation can be encouraged by employing the appropriate plant hormones in accordance with known methods and the shoots transferred to rooting medium for regeneration of plants. The plants may then be used to establish repetitive generations, either from seeds or using vegetative propagation techniques.

The constructs of the present invention can be introduced into plant cells using Ti plasmids, Ri plasmids, plant virus vectors, direct DNA transformation, micro-injection, electroporation, etc. For reviews of such techniques see for example

Weissbach and Weissbach (1988) and Geierson and Corey (1988). The present invention further includes a suitable vector comprising the chimeric gene construct.

Buttery and Buzzell (1968) showed that the amount of peroxidase activity present in seed coats may vary substantially among different cultivars. The presence of a single dominant gene *Ep* causes a high seed coat peroxidase phenotype (Buzzell and Buttery, 1969). Homozygous recessive *epep* plants are ~100-fold lower in seed coat peroxidase activity. This results from a reduction in the amount of peroxidase enzyme present, primarily in the hourglass cells of the subepidermis (Gijzen *et al.*, 1993). In plants carrying the *Ep* gene, peroxidase is heavily concentrated in the hourglass cells (osteosclereids). These cells form a highly differentiated cell layer with thick, elongated secondary walls and large intercellular spaces (Baker *et al.*, 1987).

Screening a seed coat cDNA library prepared from *EpEp* plants with a degenerate primer derived from the active site domain of plant peroxidase resulted in a high frequency of positive clones. Many of these clones encode identical cDNA molecules and indicate that the corresponding mRNA is an abundant transcript in developing seed coat tissues. The sequence of the cDNA is shown in Figure 1.

Previous studies on soybean seed coat peroxidase indicated that this enzyme is heavily glycosylated and that carbohydrate contributes 18% of the mass of the apo-enzyme (Gray *et al.*, 1996). The seven potential glycosylation sites identified from the amino acid sequence of the seed cost peroxidase (Figure 1) would accommodate the

five or six N-linked glycosylation sites proposed by Gray *et al.* (1996). The heme-binding domain encompasses residues Asp161 to Phe171 and the acid-base catalysis region from Gly33 to Cys44. The two regions are highly conserved among plant peroxidases and are centred around functional histidine residues, His169 and His40. There are eight conserved cysteine residues in the mature protein that provide for four di-sulfide bridges found in other plant peroxidases and predicted from the crystal structure of peanut peroxidase (Welinder, 1992; Schuller *et al.*, 1996). Other conserved areas include residues Cys91 to Ala105 and Val119 to Leu127 that occur in or around helix D. The most divergent aspects of the seed coat peroxidase protein sequence are the carboxy- and amino-terminal regions. These sequences probably provide special targeting signals for the proper processing and delivery of the peptide chain. It is possible the carboxy-terminal extension of the seed coat peroxidase is removed at maturity, as has been shown for certain barley and horseradish peroxidases (Welinder, 1992).

The molecular mass of the enzyme has been determined by denaturing gel electrophoresis to be 37 kDa (Sessa and Anderson, 1981; Gillikin and Graham, 1991) or 43 kDa (Gijzen *et al.*, 1993). Analysis by mass spectrometry indicated a mass of 40,622 Da for the apo-enzyme and 33,250 Da after deglycosylation (Gray *et al.*, 1996). These values are in good agreement with the mass of 35,377 Da calculated from the predicted amino acid sequence for the mature apo-protein prior to glycosylation and other modifications. Huangpu et al (1995) reported an anionic seed coat peroxidase having a M_r of 30,577 Da and characterized a partial cDNA encoding this protein.

This 1031 bp cDNA contained an open reading frame of 849 bp encoding a 283 amino acid protein. There are several differences between this reported sequence and the sequence of this invention that are manifest at the amino acid level (see Figure 3 for sequence comparison). The enzyme encoded by the gene reported by Huangpu et al is different from that of this invention as the peroxidase of this invention has a M_r of 35,377 Da.

Genomic DNA blots probed with the seed coat peroxidase cDNA produced two or three hybridizing fragments of varying intensity with most restriction enzyme digestions, despite that several peroxidase isozymes are present in soybean. The results indicate that this seed coat peroxidase is present as a single gene that does not share sufficient homology with most other peroxidase genes to anneal under conditions of high stringency.

The genomic DNA sequence comprises four exons spanning bp 1533-1752 (exon I), 2383 -2574 (exon 2), 3605-3769 (exon 3) and 4033-4516 (exon 4) and three introns comprising 1752-2382 (intron 1), 2575-3604 (intron 2) and 3770-4516 (intron 3), of SEQ ID NO:2. Features of the upstream regulatory region of the genomic DNA include a TATA box centred on bp 1487; a cap signal 32 bp down stream centred on bp 1520. Also noted within the genomic sequence are three polyadenylation signals centred on bp 4520, 4598, 4663 and a polyadenylation site at bp 4700.

This promoter is considered seed coat specific since the peroxidase protein encoded by the Ep gene accumulates in the seed coat tissues, especially in the hourglass cells of the subepidermis, and is not expressed in other tissues, aside from a marginal expression of peroxidase in the root tissues. This is also true at the transcriptional level (see Figure 9). The DNA regulatory regions of the genomic sequence of Figure 2 are used to control the expression of the adjacent peroxidase gene in seed coat tissue. Such regulatory regions include nucleotides 1-1532. Other regions of interest include nucleotides 1752-2382, 2575-3604 and/or 3770-4032 of SEQ ID NO:2. Therefore other proteins of interest may be expressed in seed coat tissues by placing a gene capable of expressing the protein of interest under the control of the DNA regulatory elements of this invention. Genes of interest include but are not restricted to herbicide resistant genes, genes encoding viral coat proteins, or genes encoding proteins conferring biological control of pest or pathogens such as an insecticidal protein for example *B. thuringiensis* toxin. Other genes include those capable of the production of proteins that alter the taste of the seed and/or that affect the nutritive value of the soybean.

A modified DNA regulatory sequence may be obtained by introducing changes into the natural sequence. Such modifications can be done through techniques known to one of skill in the art such as site-directed mutagenesis, reducing the length of the regulatory region using endonucleases or exonucleases, increasing the length through the insertion of linkers or other sequences of interest. Reducing the size of DNA regulatory region may be achieved by removing 3' or 5' regions of the regulatory

region of the natural sequence by using a endonuclease such as BAL 31 (Sambrook et al 1989). However, any such DNA regulatory region must still function as a seed coat specific DNA regulatory region.

It may be readily determined if such modified DNA regulatory elements are capable of acting in a seed coat specific manner transforming plant cells with such regulatory elements controlling the expression of a suitable marker gene, culturing these plants and determining the expression of the marker gene within the seed coat as outlined above. One may also analyze the efficacy of DNA regulatory elements by introducing constructs comprising a DNA regulatory element of interest operably linked with an appropriate marker into seed coat tissues by using particle bombardment directed to seed coat tissue and determining the degree of expression of the regulatory region as is known to one of skill in the art.

Two tandemly arranged genes encoding anionic peroxidase expressed in stems of *Populus kitakamiensis*, *prxA3a* and *prxA4a* have been cloned and characterized (Osakabe et al, 1995). Both of these genomic sequences contained four exons and three introns and encoded proteins of 347 and 343 amino acids, respectively. The two genes encode distinct isozymes with deduced M_rs of 33.9 and 34.6 kDa. Furthermore, a 532 bp promoter derived from the peroxidase gene of *Armoracia rusticana* has also been reported (Toyobo KK, JP 4,126,088, April 27, 1992). However, a search using GenBank revealed no substantial similarity between the promoter region, or introns 1, 2 and 3 of this invention and those within the literature.

Digestion of the genomic DNA with *Bam*HI or *Sac*I revealed restriction fragment length polymorphisms that distinguished *EpEp* and *epep* genotypes. Although the *Xba*I digestion did not produce a readily detectable polymorphism, the size of the hybridizing fragment in both genotypes was ~14 kb. Thus, a 0.3 kb size difference is outside of the resolving power of the separation for fragments this large. Sequence analysis of *EpEp* and *epep* genotypes indicates that the mutant *ep* allele is missing 87 bp of sequence at the 5' end of the structural gene. This would account for the drastically reduced amounts of peroxidase enzyme present in seed coats of *epep* plants since the deletion includes the translation start codon and the entire N-terminal signal sequence. However, the 87 bp deletion cannot account for the differences observed in the RFLP analysis since the missing fragment does not include a *Bam*HI site and is much smaller than the 0.3 kb polymorphism detected in the *Sac*I digestion. Thus, other genetic rearrangements must occur in the vicinity of the *ep* locus that lead to these polymorphisms.

The results shown here indicate that the mutation causing low seed coat peroxidase activity occurs in the structural gene encoding the enzyme. This mutation is an 87 bp deletion in the 5' region of the gene encompassing the translation start site. Several different low peroxidase cultivars share a similar mutation in the same area, suggesting that the recessive *ep* alleles have a common origin or that the region is prone to spontaneous deletions or rearrangements.

Due to the industrial interest in soybean seed coat peroxidase, alternate sources for the production of this enzyme are needed. The DNA of this invention, encoding the seed coat soybean peroxidase under the control of a suitable promoter and expressed within a host of interest, can be used for the preparation of recombinant soybean seed coat peroxidase enzyme.

Soybean seed coat peroxidase has been characterized as a lignin-type peroxidase that has industrially significant properties ie: high activity and stability under acidic conditions; exhibits wide substrate specificity; equivalent catalytic properties to that of *Phanerochaete chrysosporium* ligin peroxidase (the currently preferred enzyme used for treatment of industrial waste waters (Wick 1995) but is at least 150-fold more stable; more stable than horseradish peroxidase which is also used in industrial effluent treatments and medical diagnostic kits (McEldoon *et al.*, 1995). These properties are useful within industrial applications for the degradation of natural aromatic polymers including lignin and coal (McEldoon et al, 1995), and the preferred use of soybean peroxidase, over that of horseradish peroxidase, in medical diagnostic tests as an enzyme label for antigens, antibodies, oligonucleotide probes, and within staining techniques (Wick 1995). Soybean peroxidase is also used in the deinking of printed waste paper (Johnson et al., U.S. 5,270,770; December 6, 1994) and for the biocatalytic oxidation of primary alcohols (Johnson et al., U.S. 5,391,488; February 13, 1996). Soybean peroxidase has also been used as a replacement for chlorine in the pulp and paper industry, in order to remove chlorine, phenolic or aromatic amine containing pollutants from industrial waste waters (Wick 1995), or as formaldehyde

replacement (Freiberg, 1995) for use in adhesives, abrasives, and protective coatings (e.g. varnish and resins, Wick 1995).

Furthermore, the seed coat peroxidase gene may be expressed in an organ or tissue specific manner within a plant. For example, the quality and strength of cotton fibber can be improved through the over-expression of cotton or horseradish peroxidase placed under the control of a fibre-specific promoter (Maliyakal, WO 95/08914; April 6, 1995).

Similarly, seed-specific DNA regulatory regions of this invention may be used to control expression of genes of interest such as:

i) genes encoding herbicide resistance, or

ii) biological control of insects or pathogens (e.g. B. thuringiensis), or

iii) viral coat proteins to protect against viral infections, or

iv) proteins of commercial interest (e.g. pharmaceutical), and

v) proteins that alter the nutritive value, taste, or processing of seeds within the seed coat of plants.

While this invention is described in detail with particular reference to preferred embodiments thereof, said embodiments are offered to illustrate but not to limit the invention.

EXAMPLES

Plant material

All soybean (*Glycine max* [L.] Merr) cultivars and breeding lines were from the collection at Agriculture Canada, Harrow, Ontario.

Seed Coat cDNA library Construction and Screening

High seed coat peroxidase (*EpEp*) soybean cultivar Harosoy 63 plants were grown in field plots outdoors. Pods were harvested 35 days after flowering and seeds in the mid-to-late developmental stage were excised. The average fresh mass was 250 mg per seed. Seed coats were dissected and immediately frozen in liquid nitrogen. The frozen tissue was lyophilized and total RNA extracted in 100 mM Tris-HCl pH 9.0, 20 mM EDTA, 4% (w/v) sarkosyl, 200 mM NaCl, and 16 mM DTT, and precipitated with LiCl using the standard phenol/chloroform method described by Wang and Vodkin (1994). The poly(A)$^+$ RNA was purified on oligo(dT) cellulose columns prior to cDNA synthesis, size selection, ligation into the λ ZAP Express vector, and packaging according to instructions (Stratagene). A degenerate oligonucleotide with the 5' to 3' sequence of TT(C/T)CA(C/T)GA(C/T)TG(C/T)TT(C/T)GT was 5' end labelled to high specific activity and used as a probe to isolate peroxidase cDNA clones (Sambrook *et al.*, 1989). Duplicate plaque lifts were made to nylon filters (Amersham), UV fixed, and prehybridized at 36 °C for 3 h in 6 x SSC, 20 mM Na_2HPO_4 (pH6.8),

5 x Denhardt's, 0.4 % SDS, and 500 μg/mL salmon sperm DNA. Hybridization was in the same buffer, without Denhardt's, at 36 °C for 16 h. Filters were washed quickly with several changes of 6 x SSC and 0.1 % SDS, first at room temperature and finally at 40°C, prior to autoradiography for 16 h at -70°C with an intensifying screen.

Genomic DNA Isolation, Library Construction, and DNA Blot Analysis

Soybean genomic DNA was isolated from leaves of greenhouse grown plants or from etiolated seedlings grown in vermiculite. Plant tissue was frozen in liquid nitrogen and lyophilized before extraction and purification of DNA according to the method of Dellaporta *et al.* (1983). Restriction enzyme digestion of 30 μg DNA, separation on 0.5 % agarose gels and blotting to nylon membranes followed standard protocols (Sambrook *et al.*, 1989). For construction of the genomic library, DNA purified from Harosoy 63 leaf tissue was partially digested with *Bam*HI and ligated into the λ FIX II vector (Stratagene). Gigapack XL packaging extract (Stratagene) was used to select for inserts of 9 to 22 kb. After library amplification, duplicate plaque lifts were hybridized to cDNA probe.

Blots or filter lifts were prehybridized for 2 h at 65°C in 6 x SSC, 5 x Denhardt's, 0.5 % SDS, and 100 μg/mL salmon sperm DNA. Radiolabelled cDNA probe (20 to 50 ng) was prepared using the Ready-to-Go labelling kit (Pharmacia) and ^{32}P-dCTP (Amersham). Unincorporated^{32} P-dCTP was removed by spin column chromatography before adding radiolabelled cDNA to the hybridization buffer

(identical to prehybridization buffer without Denhardt's). Hybridization was for 20 h at 65°C. Membranes were washed twice for 15 min at room temperature with 2 x SSC, 0.5 % SDS, followed by two 30 min washes at 65°C with 0.1 x SSC, 0.5 % SDS. Autoradiography was for 20 h at -70°C using an intensifying screen and X-OMAT film (Kodak).

DNA Sequencing

Sequencing of DNA was performed using dye-labelled terminators and Taq-FS DNA polymerase (Perkin-Elmer). The PCR protocol consisted of 25 cycles of a 30 sec melt at 96°C, 15 sec annealing at 50°C, and 4 min extension at 60°C. Samples were analyzed on an Applied Biosystems 373A Stretch automated DNA sequencer.

Polymerase Chain Reaction

PCR amplifications contained 1 ng template DNA, 5 pmol each primer, 1.5 mM $MgCl_2$, 0.15 mM deoxynucleotide triphosphates mix, 10 mM Tris-HCl, 50 mM KCl, pH 8.3, and 1 unit of Taq polymerase (Gibco BRL) in a total volume of 25 μL. Reactions were performed in a Perkin-Elmer 480 thermal cycler. After an initial 2 min denaturation at 94°C, there were 35 cycles of 1 min denaturation at 94°C, 1 min annealing at 52°C, and 2 min extension at 72°C. A final 7 min extension at 72°C completed the program. The following primers were used for PCR analysis of genomic DNA:

prx2+	CTTCCAAATATCAACTCAAT
prx6-	TAAAGTTGGAAAAGAAAGTA
prx9	ATGCATGCAGGTTTTTCAGT
prx10-	TTGCTCGCTTTCTATTGTAT
prx12+	TCTTCGATGCTTCTTTCACC
prx29+	CATAAACAATACGTACGTGAT

RNA Isolation

For isolation of RNA, tissue was harvested from greenhouse grown plants, dissected, frozen in liquid nitrogen, and lyophilized prior to extraction. Total RNA was purified from seed coats, embryos, pods, leaves, and flowers using standard phenol/chloroform method (Sambrook et al., 1989). This method did not afford good yields of RNA from roots, therefore this tissue was extracted with Triazole reagent (GibcoBRL) and total RNA purified according to manufacturers' instructions with an additional phenol-chloroform extraction step. The amount of RNA was estimated by measuring absorbance at 260 and 280 nm, and by electrophoretic separation in formaldehyde gels followed by staining with ethidium bromide and comparison to known standards. Total RNA (10 μg per sample) was prepared, subject to electrophoresis through a 1% agarose gel containing formaldehyde, and then stained with ethidium bromide to ensure equal loading of samples. The gel was blotted to nylon (Hybond™N, Amersham) according to standard methods and the RNA was fixed to the membrane by UV cross linking.

Seed Coat Peroxidase Assays

The F_3 seed was measured for peroxidase activity to score the phenotype of the F_2 population because the seed testa is derived from maternal tissue. The seeds were briefly soaked in water and the seed coat was dissected from the embryo and placed in a vial. Ten drops (~500 μL) of 0.5% guaiacol was added and the sample was left to stand for 10 min before adding one drop (~50 μL) of 0.1% H_2O_2. An immediate change in colour of the solution, from clear to red, indicates a positive result and high seed coat peroxidase activity.

Example 1: *The Seed Coat Peroxidase cDNA and genomic DNA sequences*

To isolate the seed coat peroxidase transcript, a cDNA library was constructed from developing seed coat tissue of the *EpEp* cultivar Harosoy 63. The primary library contained 10^6 recombinant plaque forming units and was amplified prior to screening. A degenerate 17-mer oligonucleotide corresponding to the conserved active site domain of plant peroxidases was used to probe the library. In screening 10,000 plaque forming units, 12 positive clones were identified. The cDNA insert size of the clones ranged from 0.5 to 2.5 kb, but six clones shared a common insert size of 1.3 kb. These six clones (*soyprx03, soyprx05, soyprx06, soyprx11, soyprx12,* and *soyprx14*) were chosen for further characterization since the 1.3 kb insert size matched the expected peroxidase transcript size. Sequence analysis of the six clones showed that they contained identical cDNA transcripts encoding a peroxidase and that each resulted

from an independent cloning event since the junction between the cloning vector and the transcript was different in all cases.

Since it was not clear that the entire 5' end of the cDNA transcript was complete in any of the cDNA clones isolated, the structural gene corresponding to the seed coat peroxidase was isolated from a Harosoy 63 genomic library. A partial *Bam*HI digest of genomic DNA was used to construct the library and more than 10^6 plaque forming units were screened using the cDNA probe. A positive clone, G25-2-1-1-1, containing a 17 kb insert was identified and a 4.7 kb region encoding the peroxidase was sequenced SEQ ID NO:2. This region includes 1532 nucleotides of the 5' region of the peroxidase gene.

The genomic sequence matched the cDNA sequence except for three introns encoded within the gene. The genomic sequence also revealed two additional translation start codons, beginning one bp and 10 bp upstream from the 5' end of the longest cDNA transcript isolated. Figure 1 shows the deduced cDNA sequence. The open reading frame of 1056 bp encodes a 352 amino acid protein of 38,106 Da. A heme-binding domain, a peroxidase active site signature sequence, and seven potential N-glycosylation sites were identified from the deduced amino acid sequence. The first 26 amino acid residues conform to a membrane spanning domain. Cleavage of this putative signal sequence releases a mature protein of 326 residues with a mass of 35,377 Da and an estimated pI of 4.4.

Relevant features of the genomic fragment (Figure 2) include four exons at bp 192-411 (exon 1; 1533-1751 of SEQ ID NO:2), 1042 -1233 (exon 2; 2383-2574 of SEQ ID NO:2), 2263-2429 (exon 3; 4033-4516 fo SEQ ID NO:2) and 2692-3174 (exon 4; 1752-2382 of SEQ ID NO:2) and three introns at bp 412-1041 (intron 1; 1752-2382 of SEQ ID NO:2), 1234-2263 (intron 2; 2575-3604 of SEQ ID NO:2) and 2430-2691 (intron 3; 3770-4032 of SEQ ID NO:2). The 1532 bp regulatory region of the genomic DNA include a TATA box centred on bp 1487 and a cap signal 32 bp down stream centred at bp 1520 of SEQ ID NO:2. Also noted within the genomic sequence are three polyadenylation signals centred on bp 4520, 4598, 4700 and a polyadenylation site at bp 4700 of SEQ ID NO:2.

Figure 3 illustrates the relationship between the soybean seed coat peroxidase and other selected plant peroxidases. The soybean sequence is most closely related to four peroxidase cDNAs isolated from alfalfa, (see Figure 3) sharing from 65 to 67% identity at the amino acid level with the alfalfa proteins (X90693, X90694, X90692, el-Turk et al 1996; L36156, Abrahams et al 1994). When compared with other plant peroxidases, soybean seed coat peroxidase exhibits from 60 to 65% identity with poplar (D30653 and D30652, Osakabe et al 1994)) and flax (L0554, Omann and Tyson 1995); 50 to 60% identity with horseradish (M37156, Fujiyama et al. 1988), tobacco (D11396, Osakabe et al 1993), and cucumber (M91373, Rasmussen et al. 1992); and 49% identity with barley (L36093, Scott-Craig et al. 1994), wheat (X85228, Baga et al 1995) and tobacco (L02124, Diaz-De-Leon et al 1993) peroxidases.

A comparison of the promoter region, 1-1532 of SEQ ID NO:2, indicates that there are no similar sequences present within the GENBANK database.

Example 2: *DNA Blot Analysis Using the Seed Coat Peroxidase cDNA Probe Reveals Restriction Fragment Length Polymorphisms Between* EpEp *and* epep *Genotypes*

Genomic DNA blots of OX347 (*EpEp*) and OX312 (*epep*) plants were hybridized with ^{32}P-labelled cDNA to estimate the copy number of the seed coat peroxidase gene and to determine if this locus is polymorphic between the two genotypes. Figure 4 shows the hybridization patterns after digestion with *Bam*HI, *Xba*I, and *Sac*I. Restriction fragment length polymorphisms are clearly visible in the *Bam*HI and *Sac*I digestions. The *Bam*HI digestion produced a strongly hybridizing 17 kb fragment and a faint 3.4 kb fragment in the *EpEp* genotype. The 3.4 kb *Bam*HI fragment is visible in the *epep* genotype but the 17 kb fragment has been replaced by a signal at >20 kb. The *Sac*I digestion resulted in detection of three fragments in *EpEp* and *epep* plants. At least two fragments were expected here since the cDNA sequence has a *Sac*I site within the open reading frame. However, the smallest and most strongly hybridizing of these fragments is 5.2 kb in *EpEp* plants and 4.9 kb in *epep* plants. Digestion with *Xba*I produced hybridizing fragments of ~14 kb and 7.8 kb for both genotypes, with the larger fragment showing a stronger signal.

Example 3: *A Deletion Mutation Occurs in the Recessive* ep *Locus*

The structural gene encoding the seed coat peroxidase is schematically illustrated in Figure 5. The 17 kb *Bam*HI fragment encompassing the gene includes 191 bp of sequence upstream from the translation start codon, three introns of 631 bp, 1030 bp, and 263 bp, and 13 kb of sequence downstream from the polyadenylation site. The arrangement of four exons and three introns and the placement of introns within the sequence is similar to that described for other plant peroxidases (Simon, 1992; Osakabe *et al.* 1995).

Primers were designed from the DNA sequence to compare *EpEp* and *epep* genotypes by PCR analysis. Figure 6 shows PCR amplification products from four different primer combinations using OX312 (*epep*) and OX347 (*EpEp*) genomic DNA as template. The primer annealing site for prx29+ begins 182 bp upstream from the ATG start codon; the remaining primer sites are shown in Figure 1. Amplification with primers prx2+ and prx6-, and with prx12+ and prx10- produced the expected products of 1.9 kb and 860 bp, respectively, regardless of the *Ep/ep* genotype of the template DNA. However, PCR amplification with primers prx9+ and prx10-, and with prx29+ and prx10- generated the expected products only when template DNA was from plants carrying the dominant *Ep* allele. When template DNA was from an *epep* genotype, no product was detected using primers prx9+ and prx10- and a smaller product was amplified with primers prx29+ and prx10-. The products resulting from amplification of OX312 or OX347 template DNA with primers prx29+ and prx10-

were directly sequenced and compared. The polymorphism is due to an 87 bp deletion occurring within this DNA fragment in OX312 plants, as shown in Figure 5. This deletion begins nine bp upstream from the translation start codon and includes 78 bp of sequence at the 5' end of the open reading frame, including the prx9+ primer annealing site.

To test whether this deletion mutation cosegregates with the seed coat peroxidase phenotype, genomic DNA from an F_2 population segregating at the *Ep* locus was amplified using primers prx9+ and prx10- and F_3 seed was tested for seed coat peroxidase activity. Figure 7 shows the results from this analysis. Of the 30 F_2 individuals tested, all 23 that were high in seed coat peroxidase activity produced the expected 860 bp PCR amplification product. The remaining seven F_2's with low seed coat peroxidase activity produced no detectable PCR amplification products.

Finally, to determine if the OX312(*epep*) and OX347(*EpEp*) breeding lines are representative of soybean cultivars that differ in seed coat peroxidase activity, several cultivars were tested by PCR analysis using primer combinations targeted to the *Ep* locus. Figure 8 shows results from this analysis of six different soybean cultivars, three each of the homozygous dominant *EpEp* and recessive *epep* genotypes. As observed with OX312 and OX347, amplification products of the expected size were produced with primers prx12+ and prx10- regardless of the genotype, whereas *epep* genotypes yielded no product with primers prx9+ and prx10- or a smaller fragment with primers prx29+ and prx10-.

Example 4 Developmental Pattern of Expression of the Ep gene

The seed coat peroxidase mRNA levels were determined by hybridizing RNA gel blots with radio labelled cDNA probe. The figure illustrates the transcript abundance in various tissues of *epep* and *EpEp* plants. The mRNA accumulated to high levels in seed coat tissues of *EpEp* plants, especially in the later stages development when whole seed fresh weight exceeded 50 mg. Low levels of transcript could also be detected in root tissues but not in the flower, embryo, pod or leaf. The transcript could also be detected in seed coat and root tissues *epep* plants but in drastically reduced amounts compared to the *EpEp* genotype. The reduced amounts of peroxidase mRNA present in seed coats of *epep* plants indicates that the transcriptional process and/or the stability of the resulting mRNA is severely affected. The *Ep* gene has a TATA box and a 5' cap signal beginning 47 bp and 15 bp, respectively, upstream from the translation start codon. The 87 bp deletion in the *ep* allele extends into the 5' cap signal and therefore could interfere with transcript processing. Regardless, any resulting transcript will not be properly translated since the AUG initiation codon and the entire amino-terminal signal sequence is deleted from the *ep* allele. Not wishing to be bound by theory, the lack of peroxidase accumulation in seed coats of *epep* plants appears to be due to at least two factors, greatly reduced transcript levels and ineffective translation, resulting from mutation of the structural gene encoding the enzyme. In summary, the results indicate that the *Ep* gene regulatory elements can drive high level expression in a tightly coordinated, tissue and developmentally specific manner.

All scientific publications and patent documents are incorporated herein by reference.

The present invention has been described with regard to preferred embodiments. However, it will be obvious to persons skilled in the art that a number of variations and modifications can be made without departing from the scope of the invention as described in the following claims.

References

Abrahams, S.L., Hayes, C.M., and Watson, J.M. (1994) Organ-specific expression of three peroxidase-encoding cDNAs from lucerne (*Medicago sativa*). GenBank Accession # L36156.

Baga, M. Chibbar, R.N., and Kartha, K.K., (1995) Molecular cloning and expression analysis of peroxidase genes from wheat. *Plant Molec. Biol.* **29,** 647-662

Baker, D.M., Minor, H.C., and Cumbie, B.G. (1987) Scanning electron microscopy examination of soybean testa development. *Can. J. Bot.* **65,** 2420-2424.

Bowles, D.J. (1990) Defense-related proteins in higher plants. *Annu. Rev. Biochem.* **59,** 873-907.

Buttery, B.R., and Buzzell, R.I. (1968) Peroxidase activity in the seeds of soybean varieties. *Crop Sci.* **8,** 722-725.

Buzzell, R.I., and Buttery, B.R. (1969) Inheritance of peroxidase activity in soybean seed coats. *Crop Sci.* **9,** 387-388.

Campa, A. (1991) Biological roles of plant peroxidases: known and potential function. In *Peroxidases in Chemistry and Biology,* Volume II (J. Everse, K.E. Everse and M.B. Grisham, eds). Boca Raton, FL: CRC Press, pp. 25-50.

Dellaporta, S.L., Wood, J., and Hicks, J.B. (1983) A plant DNA minipreparation. Version II. *Plant Mol. Biol. Rep.* **1,** 19-21.

Diaz-De-Leon, f., Klotz, K.L., and Lagrimini, M. (1993) Nucleotide Sequence of the Tobacco (*Nicotiana tabacum*) anionic peroxidase gene. *Plant Physiol.* **101,** 1117-1118.

el-Turk, J., Asemota, O., Leymarie, J., Sallaud, C., Mesnage, S., Breda, C., Buffard, D., Kondorosi, A., and Esnault, R. (1996) Nucleotide sequence of four pathogen-induced alfalfa peroxide-encoding cDNAs. *Gene* **170,** 213-216.

Freiberg B., (1995) Indiana Crop: Keeping Its Members Up with the Changing Times. *Seed Crops Indust.* March, 4-9

Fujiyama, K., Takemura, H., Shibayama, S., Kobayashi, K., Choi, J.-K., Shinmyo, A., Takano, M., Yamada, Y,. and Okada, H. (1988) Structure fo the Horseradish Peroxidase isozyme c genes. *Eur. J. Biochem.* **173,** 681-687.

Geierson and Corey (1988), *Plant Molecular Biology*, 2d Ed.

Gijzen, M., van Huystee, R., and Buzzell, R.I. (1993) Soybean seed coat peroxidase. A comparison of high-activity and low-activity genotypes. *Plant Physiol.* **103,** 1061-1066.

Gillikin, J.W., and Graham, J.S. (1991) Purification and developmental analysis of the major anionic peroxidase from the seed coat of soybean. *Plant Physiol.* **96,** 214-220.

Gray, J.S.S., Yang, B.Y., Hull, S.R., Venzke, D.P., and Montgomery, R. (1996) The glycans of soybean peroxidase. *Glycobiology* **6,** 23-32.

Lagrimini, M.L., Bradford., and Rothstein S, (1990) Peroxidase-Induced Wilting in Transgenic Tobacco. *Plant Cell* **2,** 7-18.

McEldoon, J.P., Pokora A.R., and Dordick, J.S. (1995) Lignin peroxidase-type activity of soybean peroxidase. *Enzyme Microb. Technol.* **17,** 359-365.

Moerschbacher, B.M. (1992) Plant peroxidases: involvement in response to pathogens. In *Plant Peroxidases 1980-1990: Topics and Detailed Literature on Molecular, Biochemical, and Physiological Aspects,* (C. Penel, T. Gaspar and H. Greppin, eds). Geneva: University of Geneva, pp. 91-115.

Omann, F., and Tyson, H., (1995) cDNA sequence of a peroxidase from flax (*Linum usitissimum*), GenBank Accession # L07554.

Osakabe, K., Koyama, H., Kawai, S., Katayama, Y., and Morohoshi, N. (1993) Nucleotide sequence for the genomic DNA encoding the anionic peroxidase gene from *Nicotiana tabacum*. GenBank Accession # D11396.

Osakabe, K., Koyama, H., Kawai, S., Katayama, Y., and Morohoshi, N. (1994) Molecular cloning and nucleotide sequences of two novel cDNA that encode anionic peroxidases of *Populas kitakamiensis*. GenBank Accession # D30652.

Osakabe, K., Koyama, H., Kawai, S., Katayama, Y., and Morohoshi, N. (1995) Molecular cloning of two tandemly arranged peroxidase genes from *Populus kitakamiensis* and their differential regulation in the stem. *Plant Mol. Biol.* **28,** 677-689.

Rasmussen, J.B., Smith, J.A., Williams, S., Burkhart, W., Ward, E.R., Somerville, S.C., Ryals, J., and Hammerschmidt, R. (1992) Cloning and Systemic Expression of an acidic peroxidase associated with systemic acquired resistance to disease in cucumber. GenBank Accession # M91373.

Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual,* 2nd Edn. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Schuller, D.J., Ban, N., van Huystee, R.B., McPherson, A., and Poulos, T.L. (1996) The crystal structure of peanut peroxidase. *Structure* **4**, 311-321.

Scott-Craig, J.S., Kerby, K.B., Stein, B.D,. and Sommerville, S.C. (1994) Expression of an extracellular peroxidase that is induced in barley (*Hordeum vulgare*) by the powdery mildew pathogen (*Erysiphe graminis f. sp. hordei*). GenBank Accession # L36093.

Sessa, D.J., and Anderson, R.L. (1981) Soybean peroxidases: Purification and some properties. *J. Agric. Food Chem.* **29**, 960-965.

Simon, P. (1992) Molecular cloning of plant peroxidases. In *Plant Peroxidases 1980-1990: Topics and Detailed Literature on Molecular, Biochemical, and Physiological Aspects* (C. Penel, T. Gaspar and H. Greppin, eds) Geneva: University of Geneva, pp. 47-58.

Wang, C.S., and Vodkin, L.O. (1994) Extraction of RNA from tissues containing high levels of procyanidins that bind RNA. *Plant Mol. Biol. Rep.* **12**, 132-145.

Weissbach and Weissbach, (1988) *Methods for Plant Molecular Biology*, Academy Press, New York VIII, pp. 421-463

Wick, C.B. (1995) Enzymol International Shows Promise of Novel Peroxidase, *Chem. Eng. News*, pp. 1

Welinder, K.G. (1992) Plant peroxidase structure-function relationships. In *Plant Peroxidases 1980-1990: Topics and Detailed Literature on Molecular, Biochemical, and Physiological Aspects* (C. Penel, T. Gaspar and H. Greppin, eds) Geneva: University of Geneva, pp. 1-24.

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i) APPLICANT:

(A) NAME: Mark Gijzen

(B) STREET: 848 Princess Avenue

(C) CITY: London

(D) STATE: Ontario

(E) COUNTRY: Canada

(F) POSTAL CODE (ZIP): N5W 3M4

(ii) TITLE OF INVENTION: Seed Coat DNA Regulatory Region and Peroxidase

(iii) NUMBER OF SEQUENCES: 2

(iv) COMPUTER READABLE FORM:

(A) MEDIUM TYPE: Floppy disk

(B) COMPUTER: IBM PC compatible

(C) OPERATING SYSTEM: PC-DOS/MS-DOS

(D) SOFTWARE: Patent In Release #1.0, Version #1.30 (EPO)

(2) INFORMATION FOR SEQ ID NO: 1:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 1244 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(iii) HYPOTHETICAL: NO

(iv) ANTI-SENSE: NO

(ix) FEATURE:

(A) NAME/KEY: CDS

(B) LOCATION:1..1056

(ix) FEATURE:

(A) NAME/KEY: sig_peptide

(B) LOCATION:1..77

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 1:

```
ATG GGT TCC ATG CGT CTA TTA GTA GTG GCA TTG TTG TGT GCA TTT GCT        48
Met Gly Ser Met Arg Leu Leu Val Val Ala Leu Leu Cys Ala Phe Ala
 1               5                  10                  15

ATG CAT GCA GGT TTT TCA GTC TCT TAT GCT CAG CTT ACT CCT ACG TTC        96
Met His Ala Gly Phe Ser Val Ser Tyr Ala Gln Leu Thr Pro Thr Phe
            20                  25                  30

TAC AGA GAA ACA TGT CCA AAT CTG TTC CCT ATT GTG TTT GGA GTA ATC       144
Tyr Arg Glu Thr Cys Pro Asn Leu Phe Pro Ile Val Phe Gly Val Ile
        35                  40                  45

TTC GAT GCT TCT TTC ACC GAT CCC CGA ATC GGG GCC AGT CTC ATG AGG       192
```

Phe Asp Ala Ser Phe Thr Asp Pro Arg Ile Gly Ala Ser Leu Met Arg
50 55 60

CTT CAT TTT CAT GAT TGC TTT GTT CAA GGT TGT GAT GGA TCA GTT TTG 240
Leu His Phe His Asp Cys Phe Val Gln Gly Cys Asp Gly Ser Val Leu
65 70 75 80

CTG AAC AAC ACT GAT ACA ATA GAA AGC GAG CAA GAT GCA CTT CCA AAT 288
Leu Asn Asn Thr Asp Thr Ile Glu Ser Glu Gln Asp Ala Leu Pro Asn
85 90 95

ATC AAC TCA ATA AGA GGA TTG GAC GTT GTC AAT GAC ATC AAG ACA GCG 336
Ile Asn Ser Ile Arg Gly Leu Asp Val Val Asn Asp Ile Lys Thr Ala
100 105 110

GTG GAA AAT AGT TGT CCA GAC ACA GTT TCT TGT GCT GAT ATT CTT GCT 384
Val Glu Asn Ser Cys Pro Asp Thr Val Ser Cys Ala Asp Ile Leu Ala
115 120 125

ATT GCA GCT GAA ATA GCT TCT GTT CTG GGA GGA GGT CCA GGA TGG CCA 432
Ile Ala Ala Glu Ile Ala Ser Val Leu Gly Gly Gly Pro Gly Trp Pro
130 135 140

GTT CCA TTA GGA AGA AGG GAC AGC TTA ACA GCA AAC CGA ACC CTT GCA 480
Val Pro Leu Gly Arg Arg Asp Ser Leu Thr Ala Asn Arg Thr Leu Ala
145 150 155 160

AAT CAA AAC CTT CCA GCA CCT TTC TTC AAC CTC ACT CAA CTT AAA GCT 528
Asn Gln Asn Leu Pro Ala Pro Phe Phe Asn Leu Thr Gln Leu Lys Ala
165 170 175

TCC TTT GCT GTT CAA GGT CTC AAC ACC CTT GAT TTA GTT ACA CTC TCA 576
Ser Phe Ala Val Gln Gly Leu Asn Thr Leu Asp Leu Val Thr Leu Ser
180 185 190

GGT GGT CAT ACG TTT GGA AGA GCT CGG TGC AGT ACA TTC ATA AAC CGA 624
Gly Gly His Thr Phe Gly Arg Ala Arg Cys Ser Thr Phe Ile Asn Arg
195 200 205

TTA TAC AAC TTC AGC AAC ACT GGA AAC CCT GAT CCA ACT CTG AAC ACA 672
Leu Tyr Asn Phe Ser Asn Thr Gly Asn Pro Asp Pro Thr Leu Asn Thr
210 215 220

ACA TAC TTA GAA GTA TTG CGT GCA AGA TGC CCC CAG AAT GCA ACT GGG 720
Thr Tyr Leu Glu Val Leu Arg Ala Arg Cys Pro Gln Asn Ala Thr Gly
225 230 235 240

GAT AAC CTC ACC AAT TTG GAC CTG AGC ACA CCT GAT CAA TTT GAC AAC 768
Asp Asn Leu Thr Asn Leu Asp Leu Ser Thr Pro Asp Gln Phe Asp Asn
245 250 255

AGA TAC TAC TCC AAT CTT CTG CAG CTC AAT GGC TTA CTT CAG AGT GAC 816
Arg Tyr Tyr Ser Asn Leu Leu Gln Leu Asn Gly Leu Leu Gln Ser Asp
260 265 270

CAA GAA CTT TTC TCC ACT CCT GGT GCT GAT ACC ATT CCC ATT GTC AAT 864
Gln Glu Leu Phe Ser Thr Pro Gly Ala Asp Thr Ile Pro Ile Val Asn
275 280 285

AGC TTC AGC AGT AAC CAG AAT ACT TTC TTT TCC AAC TTT AGA GTT TCA 912
Ser Phe Ser Ser Asn Gln Asn Thr Phe Phe Ser Asn Phe Arg Val Ser
290 295 300

ATG ATA AAA ATG GGT AAT ATT GGA GTG CTG ACT GGG GAT GAA GGA GAA 960
Met Ile Lys Met Gly Asn Ile Gly Val Leu Thr Gly Asp Glu Gly Glu
305 310 315 320

ATT CGC TTG CAA TGT AAT TTT GTG AAT GGA GAC TCG TTT GGA TTA GCT 1008
Ile Arg Leu Gln Cys Asn Phe Val Asn Gly Asp Ser Phe Gly Leu Ala
325 330 335

AGT GTG GCG TCC AAA GAT GCT AAA CAA AAG CTT GTT GCT CAA TCT AAA 1056
Ser Val Ala Ser Lys Asp Ala Lys Gln Lys Leu Val Ala Gln Ser Lys
340 345 350

TAAACCAATA ATTAATGGGG ATGTGCATGC TAGCTAGCAT GTAAAGGCAA ATTAGGTTGT 1116

AAACCTCTTT GCTAGCTATA TTGAAATAAA CCAAAGGAGT AGTGTGCATG TCAATTCGAT 1176

TTTGCCATGT ACCTCTTGGA ATATTATGTA ATAATTATTT GAATCTCTTT AAGGTACTTA 1236

ATTAATCA 1244

(2) INFORMATION FOR SEQ ID NO: 2:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 4700 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: promoter

(B) LOCATION:1..1532

(ix) FEATURE:

(A) NAME/KEY: sig_peptide

(B) LOCATION:1533..1609

(ix) FEATURE:

(A) NAME/KEY: exon

(B) LOCATION:1533..1751

(ix) FEATURE:

(A) NAME/KEY: exon

(B) LOCATION:2383..2574

(ix) FEATURE:

(A) NAME/KEY: exon

(B) LOCATION:3605..3769

(ix) FEATURE:

(A) NAME/KEY: exon

(B) LOCATION:4033..4516

(ix) FEATURE:

(A) NAME/KEY: intron

(B) LOCATION:1752..1782

(ix) FEATURE:

(A) NAME/KEY: intron

(B) LOCATION:2575..3604

(ix) FEATURE:

(A) NAME/KEY: intron

(B) LOCATION:3770..4032

(ix) FEATURE:

(A) NAME/KEY: CDS

(B) LOCATION:1533..1751

(ix) FEATURE:

(A) NAME/KEY: CDS

(B) LOCATION:2383..2574

(ix) FEATURE:

(A) NAME/KEY: CDS

(B) LOCATION:3605..3769

(ix) FEATURE:

(A) NAME/KEY: CDS

(B) LOCATION:4033..4516

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 2:

TAGATAAAAA AATGGGATAT AATTTTTCTC AGATGTTGTT TATACTGTTT TTTTAATCAG 60

AATTAAAATT CCTCTTTAAT TATCGACATA ATTTTTTTTG GTGAATATTA TCGACATAAT 120

TATTTAATAC AAATTTTTAT TGTACATAGA AGTGATACTT CAATTTTAAT ATTGGAGAAC 180

AGTACGAAAA CATAAAAAAA CTGTTATTAG AAGAAAAAAA TATATGGAAA AGGTTAGCTA 240

CATATATTAG CTAAATTAGT TGTTCTAATT GGCTATATAA ACCCTATTGT ACTCTTTGTA 300

ATCTCACCTT TTTCATTTAA ATACATTTCT ACTTTTTAAG TTCTATATTT TCTCTCAATT 360

TTCTTCGATA AACCATGAAA TTTAACATGG TATATCAGCG ATACCACCCA CTTTGAAAGC 420

CATGTATGGC TAGTATGGGC AGCCAAAATT TGCCCTGGTT CAAGCAAAGC AAGTGTTTAT 480

ATAGATGTGA CTTTTGTTGA GGAACTCATG CCAATGGTAC TGATTGTGAA ACTGAGAAAA 540

CTAATTTGGA GAATTTGAAT TATGATCATT AAATACTCCT CTCCTGACTA CCTTCGTCCC 600

TCAAATTTGT ACCATCATTA TTTCCCAAAA ATTTGATTAC AATGCACTAA TTAATGAATG 660

TTTCTTACAT TATCATATTA TCATATCTGA CATTTTGTTT TTACTTTTTA TAATAATTAT 720

TTTAAAAAGT CATACATGCA AATAATTTTT TAATAGTTTA CAGTTAAATT TTTACAGTAA 780

AAATGCATGA AAATTAAACT TTATTTTTCC AAGTCATCAT TTAGTCAAAT CCCAAAACAA 840

TGATTATTTT TTGCAAATGA ATGTTTATTG AACATTTAAA TGTAGCCTAA TTAATTCTGG 900

TTATGGTGTC AATGTTCCAA AACCTAATGC AAGATCTTAG CAAGTACATA CATAGATCTA 960

ATTTTAAACT TATCTTTACG CAAGAGATAT AAAGATTATA CATCTAGTTT TAAACATTAA 1020

CTTTTGTTTT TGTGTTAAAA AACAGTAACA TTTTCTTAAT TTTGTAGAGT GACGTGCTCC 1080

AACCATATTA ACGAAGATTT TAATTGGTAT TCAAGTTCAT GAACTTAGTA AATAAGTTTT 1140

GGTCTTCAGT TTTCAATTTT CATTACAACA TTTATGTAAA ATATCAACGT TTTCTGAAAT 1200

TTGTTGCTTG TGTGCTCCAA CCACATTTAA GAGATTATAG AAATTAATTT TCAAGAAGAT 1260

AATGATTCCT ACTCTTGCTG GCCCTACCAT AGTACAATAA ATCCACTCAT AAATCAACAA 1320

GTCGTCGTCA TAGGCAATTG GGCATCATAT CATAAACAAT ACGTACGTGA TATTATCTAG 1380

TGTCTCTCAG TTTACTTTAT GAGAAATTAT TTTTCTTTAA AAAAAGTTAA TTAATAAAAA 1440

CATTTGCGAT ACCGTGAGTT ACAAGAAATC CGCCGAATTC ATCTCTATAA ATAAAAGGAT 1500

CTATATGAGA GGTAAAATCA TATTAACTCA AA ATG GGT TCC ATG CGT CTA TTA 1553
Met Gly Ser Met Arg Leu Leu
355

GTA GTG GCA TTG TTG TGT GCA TTT GCT ATG CAT GCA GGT TTT TCA GTC 1601
Val Val Ala Leu Leu Cys Ala Phe Ala Met His Ala Gly Phe Ser Val
360 365 370 375

TCT TAT GCT CAG CTT ACT CCT ACG TTC TAC AGA GAA ACA TGT CCA AAT 1649
Ser Tyr Ala Gln Leu Thr Pro Thr Phe Tyr Arg Glu Thr Cys Pro Asn
380 385 390

CTG TTC CCT ATT GTG TTT GGA GTA ATC TTC GAT GCT TCT TTC ACC GAT 1697
Leu Phe Pro Ile Val Phe Gly Val Ile Phe Asp Ala Ser Phe Thr Asp
395 400 405

CCC CGA ATC GGG GCC AGT CTC ATG AGG CTT CAT TTT CAT GAT TGC TTT 1745
Pro Arg Ile Gly Ala Ser Leu Met Arg Leu His Phe His Asp Cys Phe
410 415 420

GTT CAA GTACGTACTT TTTTTTTTCC TTCCAAAATG CCCTGCATAT TTAACAAGAT 1801
Val Gln
425

TGCTTTGTTC ACCTAGAAAA ATGTGTTTTT TTCAACGATC TTACGTACGT TTGTTTGGTT 1861

TGAAAAATAA ATCAGAAAGA GATCAAGAAA ATAGCTAGAA AGAAAGCAAC GTTTTTTTAA 1921

AAGGTATTTA GTGTGAGAAA AATATTAAAA CTGAAGAGAA AGAAATTAAA TAAGCTTTTC 1981

TTGAATGATA TTTACATGTC TTATTAACTT AAAGTCACCT TTTTTCTTTA AGTTGTGCTT 2041

GAAGAAAAAA GATGTCTTTC AGTTTAGTTT TGATTAATGC TAATTATATT TTTAATTAAT 2101

TAATTAATAC TATATATCTA TTTACCATAT TAATTATTAC TATATTTCAT GATGACAACA 2161

GACAAGTATT CTAAAGAGGT ATCGGTAGAT GATTAATTTT TTTATAAAAA AATCTTTTGC 2221

GTGTATAGAT ATTCTTTTAT AATTGGTGCA GAAACTTGTA ATGCTAATTG CAATTAATCT 2281

TACATTGATT AACTAATAGC TATAATCAAT ATTTAGGTTA GGTATAGGAG ACAAATCAAG 2341

TGATCTGAAC AAATTAAGTT GTTATATTTG CATTGTGACA G GGT TGT GAT GGA 2394
Gly Cys Asp Gly
1

TCA GTT TTG CTG AAC AAC ACT GAT ACA ATA GAA AGC GAG CAA GAT GCA 2442
Ser Val Leu Leu Asn Asn Thr Asp Thr Ile Glu Ser Glu Gln Asp Ala
5 10 15 20

CTT CCA AAT ATC AAC TCA ATA AGA GGA TTG GAC GTT GTC AAT GAC ATC 2490

Leu Pro Asn Ile Asn Ser Ile Arg Gly Leu Asp Val Val Asn Asp Ile
25 30 35

AAG ACA GCG GTG GAA AAT AGT TGT CCA GAC ACA GTT TCT TGT GCT GAT 2538
Lys Thr Ala Val Glu Asn Ser Cys Pro Asp Thr Val Ser Cys Ala Asp
40 45 50

ATT CTT GCT ATT GCA GCT GAA ATA GCT TCT GTT CTG GTAATTAATA 2584
Ile Leu Ala Ile Ala Ala Glu Ile Ala Ser Val Leu
55 60

ACTCCTAATT AATTCCCAAC CATTAAAAAG TTGCATGATT GGATTCAAAA TTCTATGGTA 2644

TTGGGGTTCT GATATAAATT TGTAATTAAA TTGCACTAAA AAAAATTATC ATATACTTTT 2704

AATAAAAAAA ATTTATCTAA TTTAATTTAT TATTAAAACT ATTTTTAAAA TTCAATCCTA 2764

ACTCTTTTTT AATCGGAGCA TGTAAGCTGG CACCCACCGT ATATCGTTGG AAGATGCTAT 2824

AAAACCATTT AATTAATGGA TGGAATCAGT CAAAACATTT AATTCAAAAT ACTCTTAATT 2884

GTGATTAGTA ATCATGTTCG GGCAAGTTAC GTTGTGTATA ATTAATTTGA CTTAATCAGA 2944

TAAAAAAACA AATGGACGCA AGCCGGTTGG TATAGATATC ACTGGCCTGT AGAATATGTG 3004

GTTTTTCACG TTTAAATAAA AGCTAGCTAC TATATTATAT TTAGTCTTTT TTTTTCTTAA 3064

ACCCATTTAA CGTGATTTAT TGACTGTGAA ACATGTTTCC ACACACAGGC TTAGAAACTC 3124

CTCGCAACTA ACATCTCCAA AATTTGACTA TTTATTTATG AAGATAATTC ATCTATGATG 3184

TTCAACTCTA TTATATATAT GTATCATCGC AGTATTAAGA ATTATAATAG TCAAATATAG 3244

AAGTATATCG GGTAAATGTA GTTGCATGTG CGACCTGTTT CGTGTAAAAT GCTTATTCTA 3304

TATAGCTTTT TTTATTGGAA AATAACGATG AACTAAAAAC GAAAGGGTAT CATATAGTTT 3364

GACTTTTATG TTAGAGAGAG ACATCTTAAT TTGGTCATAT GTTAAATAAT TAATTACAAT 3424

GCATACACAA ATATTTATGC CATATCTAAA AAATGATAAA ATATCATAGG TATACTCAAC 3484

TATATGATAT CCCCATAACA GAAATTGTAC TTTTCTTCAG GCAATGAACT TAACATTTCT 3544

GTTTGCTAAA AACAAACATC CACTTAAAGT GGTTCAACAT ATTTATGTAA TAATTTACAG 3604

GGA GGA GGT CCA GGA TGG CCA GTT CCA TTA GGA AGA AGG GAC AGC TTA 3652
Gly Gly Gly Pro Gly Trp Pro Val Pro Leu Gly Arg Arg Asp Ser Leu
1 5 10 15

ACA GCA AAC CGA ACC CTT GCA AAT CAA AAC CTT CCA GCA CCT TTC TTC 3700
Thr Ala Asn Arg Thr Leu Ala Asn Gln Asn Leu Pro Ala Pro Phe Phe
20 25 30

AAC CTC ACT CAA CTT AAA GCT TCC TTT GCT GTT CAA GGT CTC AAC ACC 3748
Asn Leu Thr Gln Leu Lys Ala Ser Phe Ala Val Gln Gly Leu Asn Thr
35 40 45

CTT GAT TTA GTT ACA CTC TCA GGTATACATA ATCAATTTTT TATTTGCTAT 3799
Leu Asp Leu Val Thr Leu Ser
50 55

TAGCTAGCAA TAAAAAGTCT CTGATACAGA CATATTTAGA TAAATTAATT TCTCCATAAA 3859

CATTTATAAT AAAATTATCA ATTTATGTAC TTAAAAATTA TGGATTGAAG CTCTTTTCAT 3919

CCAACTTTTA CTAAAGTTAA GGTGCATATA ATATAAAATA AACTATCTCT TGTTTCTTAT 3979

AAAAAGATTG AAGATAAGTT AAAGTCTACT TATAAATCAT TAATATATGT ATA GGT 4035
Gly
1

GGT CAT ACG TTT GGA AGA GCT CGG TGC AGT ACA TTC ATA AAC CGA TTA 4083
Gly His Thr Phe Gly Arg Ala Arg Cys Ser Thr Phe Ile Asn Arg Leu
5 10 15

TAC AAC TTC AGC AAC ACT GGA AAC CCT GAT CCA ACT CTG AAC ACA ACA 4131
Tyr Asn Phe Ser Asn Thr Gly Asn Pro Asp Pro Thr Leu Asn Thr Thr
20 25 30

TAC TTA GAA GTA TTG CGT GCA AGA TGC CCC CAG AAT GCA ACT GGG GAT 4179
Tyr Leu Glu Val Leu Arg Ala Arg Cys Pro Gln Asn Ala Thr Gly Asp
35 40 45

AAC CTC ACC AAT TTG GAC CTG AGC ACA CCT GAT CAA TTT GAC AAC AGA 4227
Asn Leu Thr Asn Leu Asp Leu Ser Thr Pro Asp Gln Phe Asp Asn Arg
50 55 60 65

TAC TAC TCC AAT CTT CTG CAG CTC AAT GGC TTA CTT CAG AGT GAC CAA 4275
Tyr Tyr Ser Asn Leu Leu Gln Leu Asn Gly Leu Leu Gln Ser Asp Gln
70 75 80

GAA CTT TTC TCC ACT CCT GGT GCT GAT ACC ATT CCC ATT GTC AAT AGC 4323
Glu Leu Phe Ser Thr Pro Gly Ala Asp Thr Ile Pro Ile Val Asn Ser
85 90 95

```
TTC AGC AGT AAC CAG AAT ACT TTC TTT TCC AAC TTT AGA GTT TCA ATG      4371
Phe Ser Ser Asn Gln Asn Thr Phe Phe Ser Asn Phe Arg Val Ser Met
        100                 105                 110

ATA AAA ATG GGT AAT ATT GGA GTG CTG ACT GGG GAT GAA GGA GAA ATT      4419
Ile Lys Met Gly Asn Ile Gly Val Leu Thr Gly Asp Glu Gly Glu Ile
    115                 120                 125

CGC TTG CAA TGT AAT TTT GTG AAT GGA GAC TCG TTT GGA TTA GCT AGT      4467
Arg Leu Gln Cys Asn Phe Val Asn Gly Asp Ser Phe Gly Leu Ala Ser
130                 135                 140                 145

GTG GCG TCC AAA GAT GCT AAA CAA AAG CTT GTT GCT CAA TCT AAA TAA      4515
Val Ala Ser Lys Asp Ala Lys Gln Lys Leu Val Ala Gln Ser Lys  *
                150                 155                 160

ACCAATAATT AATGGGGATG TGCATGCTAG CTAGCATGTA AAGGCAAATT AGGTTGTAAA    4575

CCTCTTTGCT AGCTATATTG AAATAAACCA AAGGAGTAGT GTGCATGTCA ATTCGATTTT    4635

GCCATGTACC TCTTGGAATA TTATGTAATA ATTATTTGAA TCTCTTTAAG GTACTTAATT    4695

AATCA                                                               4700
```

THE EMBODIMENTS OF THE INVENTION IN WHICH AN EXCLUSIVE PROPERTY OF PRIVILEGE IS CLAIMED ARE DEFINED AS FOLLOWS:

1. An isolated DNA molecule comprising the nucleotide sequence of SEQ ID NO:1.

2. An isolated DNA molecule comprising at least 24 contiguous nucleotides selected from nucleotides 1-1532 of SEQ ID NO:2

3. The isolated DNA molecule comprising a nucleotide sequence substantially homologous to nucleotides 1533-4700 of SEQ ID NO:2.

4. The isolated DNA molecule of claim 3 comprising a nucleotide sequence substantially homologous to that of nucleotides 1-4700 of SEQ ID NO:2.

5. The isolated DNA molecule of claim 3 comprising nucleotides 1533-4700 of SEQ ID NO:2.

6. The isolated DNA molecule of claim 4 comprising the nucleotide sequence of SEQ ID NO:2.

7. The isolated DNA molecule of claim 2 comprising a nucleotide sequence substantially homologous to that of 1-1532 of SEQ ID NO:2.

8. The isolated DNA molecule of claim 7, comprising the nucleotide sequence of nucleotides 1-1532 of SEQ ID NO:2.

9. An isolated DNA molecule of claim 3 comprising at least 32 contiguous nucleotides selected from nucleotides 412-1041 of SEQ ID NO:2.

10. An isolated DNA molecule of claim 9 comprising the nucleotide sequence of 412-1041 of SEQ ID NO:2.

11. An isolated DNA molecule of claim 3 comprising at least 23 contiguous nucleotides selected from nucleotides 1234-2263 of SEQ ID NO:2.

12. An isolated DNA molecule of claim 11 comprising the nucleotide sequence of 1234-2263 of SEQ ID NO:2.

13. An isolated DNA molecule of claim 3 comprising at least 22 contiguous nucleotides selected from nucleotides 2430-2691 of SEQ ID NO:2.

14. An isolated DNA molecule of claim 13 comprising the nucleotide sequence of 2430-2691 of SEQ ID NO:2.

15. A vector which comprises the DNA molecule of claim 1.

16. A vector which comprises the DNA molecule of claim 2.

17. A vector which comprises the DNA molecule of claim 3.

18. The vector of claim 16 which comprises a heterologous gene of interest under control of the DNA molecule.

19. A host cell capable of expressing the DNA molecule within the vector of claim 15.

20. A host cell capable of expressing the DNA molecule within the vector of claim 16.

21. A host cell capable of expressing the DNA molecule within the vector of claim 17.

22. A host cell capable of expressing the DNA molecule within the vector of claim 18.

23. A transgenic plant comprising the vector of claim 15.

24. A transgenic plant comprising the vector of claim 16.

25. A transgenic plant comprising the vector of claim 17.

26. A transgenic plant comprising the vector of claim 18.

27. A method for the production of soybean seed coat peroxidase in a host cell comprising:

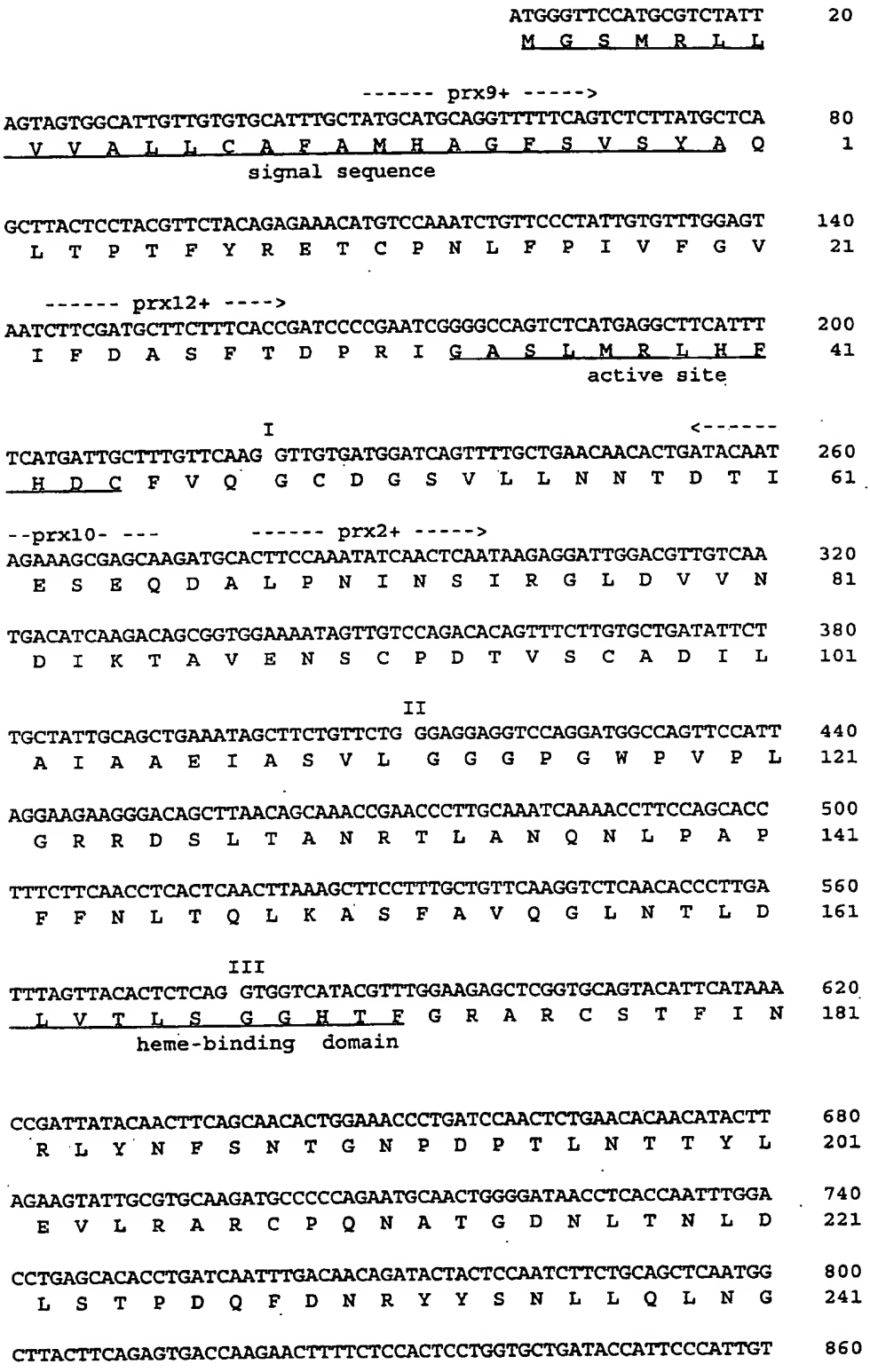
i) transforming the host cell with a vector comprising an isolated DNA molecule selected from the group consisting of SEQ ID NO:1, and SEQ ID NO:2; and

ii) culturing the host cell under conditions to allow expression of the soybean seed coat peroxidase.

28. A process for producing a heterologous gene of interest comprising propagating a transformed plant with the vector of claim 16.

29. The process of claim 28 wherein the heterologous gene of interest is produced within seed coat cells.

FIGURE 1

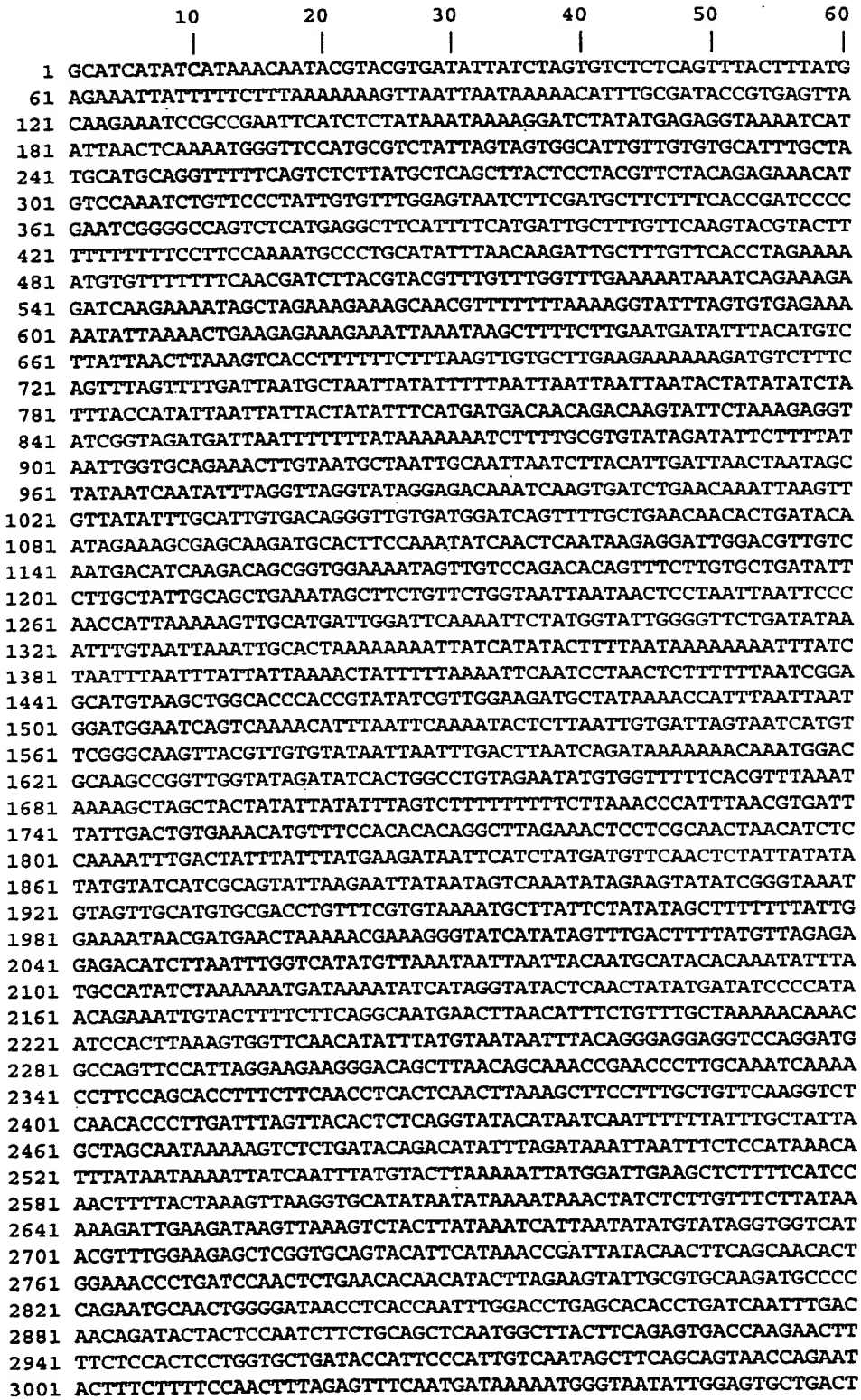


Gowling, Strathy & Henderson

```
  L  L  Q  S  D  Q  E  L  F  S  T  P  G  A  D  T  I  P  I  V    261

                              <------ prx6- ------
CAATAGCTTCAGCAGTAACCAGAATACTTTCTTTTCCAACTTTAGAGTTTCAATGATAAA    920
  N  S  F  S  S  N  Q  N  T  F  F  S  N  F  R  V  S  M  I  K    281

AATGGGTAATATTGGAGTGCTGACTGGGGATGAAGGAGAAATTCGCTTGCAATGTAATTT    980
  M  G  N  I  G  V  L  T  G  D  E  G  E  I  R  L  Q  C  N  F    301

TGTGAATGGAGACTCGTTTGGATTAGCTAGTGTGGCGTCCAAAGATGCTAAACAAAAGCT   1040
  V  N  G  D  S  F  G  L  A  S  V  A  S  K  D  A  K  Q  K  L    321

TGTTGCTCAATCTAAATAAACCAATAATTAATGGGGATGTGCATGCTAGCTAGCATGTAA   1100
  V  A  Q  S  K  *                                              326

AGGCAAATTAGGTTGTAAACCTCTTTGCTAGCTATATTGAAATAAACCAAAGGAGTAGTG   1160

TGCATGTCAATTCGATTTTGCCATGTACCTCTTGGAATATTATGTAATAATTATTTGAAT   1220

CTCTTTAAGGTACTTAATTAATC(A)n
```

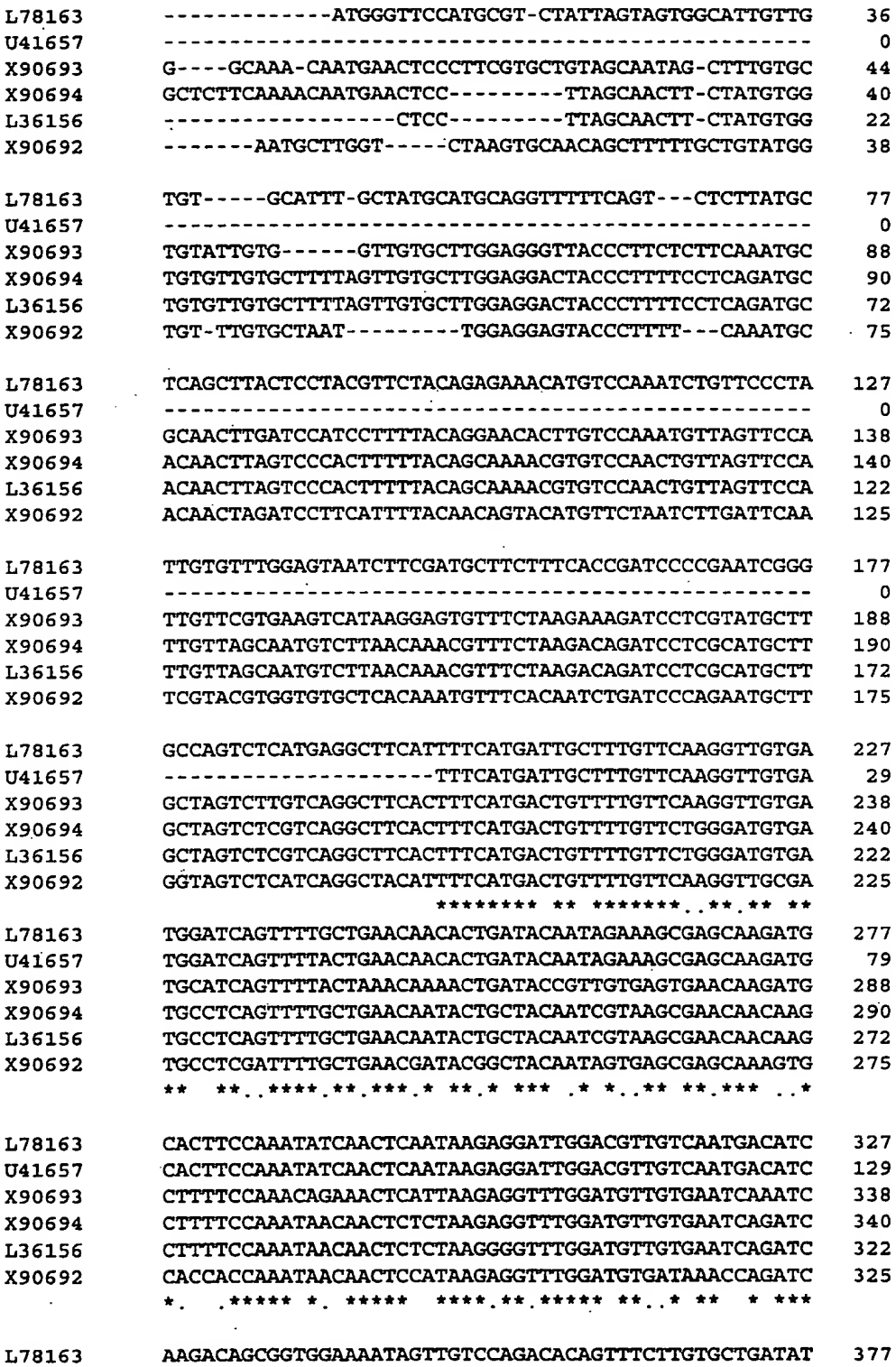
Gowling, Strathy & Henderson

FIGURE 2


10 20 30 40 50 60
1 GCATCATATCATAAACAATACGTACGTGATATTATCTAGTGTCTCTCAGTTTACTTTATG
61 AGAAATTATTTTTCTTTAAAAAAAGTTAATTAATAAAAACATTTGCGATACCGTGAGTTA
121 CAAGAAATCCGCCGAATTCATCTCTATAAATAAAAGGATCTATATGAGAGGTAAAATCAT
181 ATTAACTCAAAATGGGTTCCATGCGTCTATTAGTAGTGGCATTGTTGTGTGCATTTGCTA
241 TGCATGCAGGTTTTTCAGTCTCTTATGCTCAGCTTACTCCTACGTTCTACAGAGAAACAT
301 GTCCAAATCTGTTCCCTATTGTGTTTGGAGTAATCTTCGATGCTTCTTTCACCGATCCCC
361 GAATCGGGGCCAGTCTCATGAGGCTTCATTTTCATGATTGCTTTGTTCAAGTACGTACTT
421 TTTTTTTTCCTTCCAAAATGCCCTGCATATTTAACAAGATTGCTTTGTTCACCTAGAAAA
481 ATGTGTTTTTTTCAACGATCTTACGTACGTTTGTTTGGTTTGAAAAATAAATCAGAAAGA
541 GATCAAGAAAATAGCTAGAAAGAAAGCAACGTTTTTTTAAAAGGTATTTAGTGTGAGAAA
601 AATATTAAAACTGAAGAGAAAGAAATTAAATAAGCTTTTCTTGAATGATATTTACATGTC
661 TTATTAACTTAAAGTCACCTTTTTTCTTTAAGTTGTGCTTGAAGAAAAAAGATGTCTTTC
721 AGTTTAGTTTTGATTAATGCTAATTATATTTTTAATTAATTAATTAATACTATATATCTA
781 TTTACCATATTAATTATTACTATATTTCATGATGACAACAGACAAGTATTCTAAAGAGGT
841 ATCGGTAGATGATTAATTTTTTTTATAAAAAAAATCTTTTGCGTGTATAGATATTCTTTTAT
901 AATTGGTGCAGAAACTTGTAATGCTAATTGCAATTAATCTTACATTGATTAACTAATAGC
961 TATAATCAATATTTAGGTTAGGTATAGGAGACAAATCAAGTGATCTGAACAAATTAAGTT
1021 GTTATATTTGCATTGTGACAGGGTTGTGATGGATCAGTTTTGCTGAACAACACTGATACA
1081 ATAGAAAGCGAGCAAGATGCACTTCCAAATATCAACTCAATAAGAGGATTGGACGTTGTC
1141 AATGACATCAAGACAGCGGTGGAAAATAGTTGTCCAGACACAGTTTCTTGTGCTGATATT
1201 CTTGCTATTGCAGCTGAAATAGCTTCTGTTCTGGTAATTAATAACTCCTAATTAATTCCC
1261 AACCATTAAAAAGTTGCATGATTGGATTCAAAATTCTATGGTATTGGGGTTCTGATATAA
1321 ATTTGTAATTAAATTGCACTAAAAAAAATTATCATATACTTTTAATAAAAAAAATTTATC
1381 TAATTTAATTTATTATTAAAACTATTTTTAAAATTCAATCCTAACTCTTTTTTAATCGGA
1441 GCATGTAAGCTGGCACCCACCGTATATCGTTGGAAGATGCTATAAAACCATTTAATTAAT
1501 GGATGGAATCAGTCAAAACATTTAATTCAAAATACTCTTAATTGTGATTAGTAATCATGT
1561 TCGGGCAAGTTACGTTGTGTATAATTAATTTGACTTAATCAGATAAAAAAACAAATGGAC
1621 GCAAGCCGGTTGGTATAGATATCACTGGCCTGTAGAATATGTGGTTTTTCACGTTTAAAT
1681 AAAAGCTAGCTACTATATTATATTTAGTCTTTTTTTTTCTTAAACCCATTTAACGTGATT
1741 TATTGACTGTGAAACATGTTTCCACACACAGGCTTAGAAACTCCTCGCAACTAACATCTC
1801 CAAAATTTGACTATTTATTTATGAAGATAATTCATCTATGATGTTCAACTCTATTATATA
1861 TATGTATCATCGCAGTATTAAGAATTATAATAGTCAAATATAGAAGTATATCGGGTAAAT
1921 GTAGTTGCATGTGCGACCTGTTTCGTGTAAAATGCTTATTCTATATAGCTTTTTTTATTG
1981 GAAAATAACGATGAACTAAAAACGAAAGGGTATCATATAGTTTGACTTTTATGTTAGAGA
2041 GAGACATCTTAATTTGGTCATATGTTAAATAATTAATTACAATGCATACACAAATATTTA
2101 TGCCATATCTAAAAAATGATAAAATATCATAGGTATACTCAACTATATGATATCCCCATA
2161 ACAGAAATTGTACTTTTCTTCAGGCAATGAACTTAACATTTCTGTTTGCTAAAAACAAAC
2221 ATCCACTTAAAGTGGTTCAACATATTTATGTAATAATTTACAGGGAGGAGGTCCAGGATG
2281 GCCAGTTCCATTAGGAAGAAGGGACAGCTTAACAGCAAACCGAACCCTTGCAAATCAAAA
2341 CCTTCCAGCACCTTTCTTCAACCTCACTCAACTTAAAGCTTCCTTTGCTGTTCAAGGTCT
2401 CAACACCCTTGATTTAGTTACACTCTCAGGTATACATAATCAATTTTTTATTTGCTATTA
2461 GCTAGCAATAAAAAGTCTCTGATACAGACATATTTAGATAAATTAATTTCTCCATAAACA
2521 TTTATAATAAAATTATCAATTTATGTACTTAAAAATTATGGATTGAAGCTCTTTTCATCC
2581 AACTTTTACTAAAGTTAAGGTGCATATAATATAAAATAAACTATCTCTTGTTTCTTATAA
2641 AAAGATTGAAGATAAGTTAAAGTCTACTTATAAATCATTAATATATGTATAGGTGGTCAT
2701 ACGTTTGGAAGAGCTCGGTGCAGTACATTCATAAACCGATTATACAACTTCAGCAACACT
2761 GGAAACCCTGATCCAACTCTGAACACAACATACTTAGAAGTATTGCGTGCAAGATGCCCC
2821 CAGAATGCAACTGGGGATAACCTCACCAATTTGGACCTGAGCACACCTGATCAATTTGAC
2881 AACAGATACTACTCCAATCTTCTGCAGCTCAATGGCTTACTTCAGAGTGACCAAGAACTT
2941 TTCTCCACTCCTGGTGCTGATACCATTCCCATTGTCAATAGCTTCAGCAGTAACCAGAAT
3001 ACTTTCTTTTCCAACTTTAGAGTTTCAATGATAAAAATGGGTAATATTGGAGTGCTGACT

```
3061 GGGGATGAAGGAGAAATTCGCTTGCAATGTAATTTTGTGAATGGAGACTCGTTTGGATTA
3121 GCTAGTGTGGCGTCCAAAGATGCTAAACAAAAGCTTGTTGCTCAATCTAAATAAACCAAT
3181 AATTAATGGGGATGTGCATGCTAGCTAGCATGTAAAGGCAAATTAGGTTGTAAACCTCTT
3241 TGCTAGCTATATTGAAATAAACCAAAGGAGTAGTGTGCATGTCAATTCGATTTTGCCATG
3301 TACCTCTTGGAATATTATGTAATAATTATTTGAATCTCTTTAAGGTACTTAATTAATCA
```

FIGURE 3A



Gowling, Strathy & Henderson

```
U41657      AAGACAGCGGTGGAAAATAGTTGTCCAGACACAGTTTCTTGTGCTGATAT  179
X90693      AAAACAGCTGTGGAAAAGGCTTGTCCTAACACAGTTTCTTGTGCTGATAT  388
X90694      AAACTGGCTGTAGAAGTGCCTTGTCCTAACACAGTTTCTTGTGCTGATAT  390
L36156      AAAACTGCTGTAGAAAGTGCTTGTCCTAACACAGTTTCTTGTGCTGATAT  372
X90692      AAAACAGCGGTGGAAAATGCTTGTCCTAACACAGTTTCTTGTGCTGATAT  375
            **.  .**.**.***...  ******..**********************

L78163      TCTTGCTATTGCAGCTGAAATAGCTTCTGTT-CTGGGAGGAGGTCCAGGA  426
U41657      TCTTGCTATTGCAGCTGAAATAGCTTCTGTTGCTGGGAGGAGGTC-AGGA  228
X90693      TCTTGCTCTTTCTGCTGAATTATCATCTACA-CTGGCAGATGGTCCTGAC  437
X90694      TCTTGCACTTGCTGCTCAAGCATCCTCTGTT-CTGGCACAAGGTCCTAGT  439
L36156      TCTTGCACTTGCT---CAAGCATCCTCTGTT-CTGGCACAAGGTCCTAGT  418
X90692      TCTTGCTCTTTCTGCTGAAATATCATCTGAT-CTGGCAAATGGTCCTACT  424
            ******. **.*.    **. *.* ***. . **** * ..**** ..

L78163      TGGCCAGTTCCATTAGGAAGAAGGGACAGCTTAACAGCAAACCGAACCCT  476
U41657      TGGCCAGTTCCATTAGGAAGAAGGGACAGCTTAACAGCAAACCGAACCCT  278
X90693      TGGAAGGTTCCTTTAGGAAGAAGAGATGGTTTAACGGCAAACCAGTTACT  487
X90694      TGGACGGTTCCTTTAGGAAGAAGGGATGGTTTAACCGCAAACCGAACACT  489
L36156      TGGACGGTTCCTTTAGGAAGAAGGGATGGTTTAACCGCAAACCGAACACT  468
X90692      TGGCAAGTTCCATTAGGAAGAAGGGATAGTTTGACAGCAAATAATTCCCT  474
            ***  .*****.***********.** .* **.** *****  ...  **

L78163      TGCAAATCAAAACCTTCCAGCACCTTTCTTCAA--CCTCA-CTCAACTTA  523
U41657      TGCAAATCAAAACCTTCCAGCACCTTTCTTCAA--CCTCA-CTCAACTTA  325
X90693      TGCTAATCAAAATCTTCCAGCTCC---TTTCAATACTACTGATCAACTTA  534
X90694      TGCAAATCAAAATCTTCCGGCTCC---ATTCAATTCCTTGGATCAACTTA  536
L36156      TGCAAATCAAAATCTTCCGGCTCC---ATTCAATTCCTTGGATCACCTTA  515
X90692      TGCAGCTCAAAATCTTCCTGCCCCCACTTTCAA--CCTTA-CTCGACTAA  521
            ***.. ****** *****.** **    *****  * . .  **. **.*

L78163      AAGCTTCCTTTG-CTGTTCAAGGTCTCAACACCCTTGATTTAGTTACACT  572
U41657      AAGCTTCCTTTG-CTGTTCAAGGTCTCAACACCCTTGATTTAGTTACACT  374
X90693      AAGCTGCATTTG-CTGCTCAAGGTCTCGATACTACTGATCTGGTTGCACT  583
X90694      AAGCTGCATTT-ACTGCTCAAGGCCTCAATACTACTGATCTAGTTGCACT  585
L36156      AA-CTGCATTTGACTGCTCAAGGCCTCATTACTCCTGTTCTAGTTGCCCT  564
X90692      AATCTAACTTTGA-TAATCAAAACCTCAGTACTACTGATCTAGTTGCACT  570
            ** **.  ***   *. ****.. ***.. **   **.* *.***.* **

L78163      CTCAGGTGGTCATACGTTTGGAAGAGCTCGGTGCAGTACATTCATAAACC  622
U41657      CTCAGGTGGTCATACGTCTGGAAGAGCTCGGTGCAGTACATTCATAAACC  424
X90693      CTCCGGTGCTCATACATTTGGAAGAGCTCATTGCTCTTTATTTGTTAGCC  633
X90694      CTCGGGTGCTCATACATTTGGAAGAGCTCATTGCGCACAATTTGTTAGTC  635
L36156      CTCGGGTGCTCATACATTTGGAAGAGCTCATTGCGCACAATTTGTTAGTC  614
X90692      CTCAGGTGGCCATACAATTGGAAGAGGTCAATGCAGATTTTTCGTTGATC  620
            *** ****  *****.. ******** **..***. . . .** .*... *

L78163      GATTATACAACTTCAGCAACACTGGAAACCCTGATCCAACTCTGAACACA  672
U41657      GATTATACAACTTCAGCAACACTGGA----CTGATCCA-CT-TGGACACA  468
```

```
X90693      GATTGTACAACTTCAGCGGTACGGGAAGTCCCGATCCAACTCTTAACACA  683
X90694      GATTGTACAACTTCAGCAGTACTGGAAGTCCCGATCCAACTCTTAACACA  685
L36156      GATTGTACAACTTCAGCAGTACTGGAAGTCCCGATCCAACTCTTAACACA  664
X90692      GATTATACAATTTCAGCAACACTGGAAACCCCGATTCAACTCTTAACACG  670
            ****.***** ******.. **.***    * *** ** ** *..****.

L78163      ACATACTTAGAAGTATTGCGTGCAAGATGCCCCCAGAATGCAACTGGGGA  722
U41657      ACATACTTAGAAGTATTGCGTGCAAGATGCCCCCAGAATGCAACTGGGGA  518
X90693      ACTTACTTACAACAATTGCGCACAATATGTCCCAATGGTGGACCTGGCAC  733
X90694      ACTTACTTACAACAACTGCGCACAATATGTCCCAATGGTGGACCTGGCAC  735
L36156      ACTTACTTACAACAACTGCGCACAATATGTCCCAATGGTGGACCTGGCAC  714
X90692      ACCTATTTACAAACATTGCAAGCAATATGTCCCAATGGTGGACCTGGTAC  720
            ** ** *** **  * ***. .***.*** *** *...** * **** .

L78163      TAACCTCACCAATTTGGACCTGAGCACACCTGATCAATTTGACAACAGAT  772
U41657      TAACCTCACCAATTTGGACCTGAGCACACCTGATCAATTTGACAACAGAT  568
X90693      GAACCTTACCAATTTCGATCCAACGACTCCTGATAAATTTGACAAGAACT  783
X90694      AAACCTTACCAATTTCGATCCAACGACTCCTGATAAATTTGACAAGAACT  785
L36156      AAACCTTACCAATTTCGATCCAACGACTCCTGATAAATTTGACAAGAACT  764
X90692      AAACCTAACCGATTTGGACCCAACCACACCAGATACATTTGACTCCAACT  770
            .***** ***.**** ** * .*  **.**.***  *******.  *. *


L78163      ACTACTCCAATCTTCTGCAGCTCAATGGCTTACTTCAGAGTGACCAAGAA  822
U41657      ACTACTCCAATCTTCTGCAGCTCAATGGCTTACTTCAGAGTGACCAAGAA  618
X90693      ATTACTCTAATCTTCAAGTGAAAAAAGGTTTGCTTCAAAGTGATCAAGAG  833
X90694      ATTACTCCAATCTTCAAGTGAAAAAGGGTTTGCTCCAAAGTGATCAAGAG  835
L36156      ATTACTCCAATCTTCAAGTGAAAAAGGGTTTGCTCCAAAGTGATCAAGAG  814
X90692      ACTACTCCAATCTCCAAGTTGGAAAGGGCTTGTTTCAGAGTGACCAAGAG  820
            * ***** ***** *.. .. . **.** **. * **.***** *****.

L78163      CTTTTCTCCACTCCTGGTGCTGATACCATTCCCATTGTCAATAGCTTCAG  872
U41657      CGTTTCTCCACTCCTGGTGCTGATACCATTCC-ATTGTCAATAGCTTCAG  667
X90693      TTGTTCTCAACATCTGGTTCAGATACCATTAGCATTGTCAACAAATTCGC  883
X90694      TTGTTCTCAACTTCTGGTGCAGATACCATTAGCATTGTCAACAAATTCAG  885
L36156      TTGTTCTCAACTTCTGGTGCAGATACCATTAGCATTGTCGACAAATTCAG  864
X90692      CTTTTTTCCAGAAATGGTTCTGACACTATTTCTATTGTCAATAGTTTCGC  870
             ..** ** * .  ****.*.** ** ***   ******.* *. ***.

L78163      CAGTAACCAGAATACTTTCTTTTCCAACTTTAGAGTTTCAATGATAAAAA  922
U41657      CG--AACCAGAATACTTTCTTTTCCAACTTTAGAGTTTCAATGATAAAAA  715
X90693      AACCGATCAAAAAGCTTTTTTTGAGAGCTTTAGGGCTGCTATGATCAAAA  933
X90694      CACCGATCAAAATGCTTTCTTTGAGAGCTTTAAGGCTGCAATGATTAAAA  935
L36156      CACCGATCAAAATGCTTTCTTTGAGAGCTTTAAGGCTGCAATGATTAAAA  914
X90692      CAATAATCAAACTCTCTTCTTTGAAAATTTTGTAGCCTCAATGATAAAAA  920
             .  .*.**.* .   ** ***.  *. ***...*  .*.***** ****

L78163      TGGGTAATATTGGAGTGCTGACTGGGGATGAAGGAGAAATTCGCTTGCAA  972
U41657      TGGGTAATATTGGAGTGCTGACTGGGGATGAAGGAGAAATTCGCTTGCAA  765
X90693      TGGGAAATATTGGTGTGTTAACCGGGAACCAAGGAGAGATTAGAAAACAA  983
X90694      TGGGCAATATTGGTGTGCTAACAGGGACAAAAGGAGAGATTAGAAAACAA  985
L36156      TGGGCAATATTGGTGTGCTAACAGGGACAAAAGGAGAGATTAGAAAACAA  964
X90692      TGGGTAATATTGGAGTTTTAACTGGATCTCAAGGTGAAATTAGAACACAG  970
```

```
                **** ********.**. *.** **..   ****.**.*** * . .**.

L78163      TGTAATTTTGTGAA---TGGAGACTCGT-----------TTGGATTAGC  1007
U41657      TGTAATTTTGTGAA---TGGAGACTCGT-----------TTGGATTAGC   800
X90693      TGCAACTTTGTTAATT---------CAAAATCAGCAGAACTTGGTCTTAT  1024
X90694      TGCAACTTTGTGAACTTTGTGAACTCAAATTCTGCAGAACTAGATTTAGC  1035
L36156      TGCAACTT---------TGTGAACTCAAATTCTGCAGAACTAGATTTAGC  1005
X90692      TG---------TAATGCTGTGAATGGGAATTCTTC------TGGATTGGC  1005
            **                              ..            .*.. *..

L78163      TAGTGTGGCGTCCAAAGATGCTAAACAAAAGCTTGTTGCTCAATCTAAAT  1057
U41657      TAGTGTGGCGTCCAAAGATGCTAAACAAAAGCTTGTTGCTCAATCTAAAT   850
X90693      CAATGTTGCCTC---AGCAG--ATTCATCTG-AGGAGGGTATGGTTAG--  1066
X90694      CACCATAGCATCCATAGTAG--AATCATTAG-AGGATGGTATTGCTAGTG  1082
L36156      CACCATAGCATCCATAGTAG--AATCATTAG-AGGATGGAATTGCTAGTG  1052
X90692      TACTGTAGTCACCAA---AG--AATCATCAG-AAGATGGAATGGCTAGCT  1049
             *  .*.*  .*      .*  *..**. .* ..*..* . ... **.

L78163      AAACCAATAATTAATGGGGATGTGCATGCTAGCTAGCATGTAAAGGCAAA  1107
U41657      AAACCAATAATTAATGGGGATGTCGATGCTAGCTACGATGTAAAGGCAAA   900
X90693      ---------------------------------CTCAATGTAAA-TG-TAG  1082
X90694      TAATATAAATAAATTAG-------CGTAAATGCACTTATTGAA-ATCTTG  1124
L36156      TAATATAAATAAATTAG-------CGAAAATGCACTTATTGAA-ATCTTG  1094
X90692      CATTCTAAAT--ATAAG-------CTTGGAAAATATTGAAGAGGTTCTAT  1090
                                          .  ....*. .. ...

L78163      TTAGGTTGTAAACCTCTTTGCTAGCTATATTGAAATAAACCAAAGGAGTA  1157
U41657      TTAGGTTG-AAACCTCTTTGCTAGCTATATTGAAATAAACCAAAGGAGTA   949
X90693      T--GATTGGAAGCAACTAA--TAAATTAAGAAGCTATAAC---------T  1119
X90694      T--GACTAGATGCCACTAA--TAAAT----AAGTTATAAC---------T  1157
L36156      T--GACTAGATCCCACTAA--TAAAT----AAGTTATAAC---------T  1127
X90692      A--ATTTTGTGCATACATA--TATGGTATGTG-----------------  1118
            .  .. *. ..   .*...  **. .    ..


L78163      GTGTGCATGTCAATTCGATTTTGC-CATGTACCTCTTGGAATAT------  1200
U41657      GTGTCGATGTCAATTCGATTTTGC-CATGTACCTCTTGGAATATTATGTA   998
X90693      ATGCACATT-CATGGTATGTGTGAGATAGTTATTAGATGCTTTGTGAGCA  1168
X90694      AGGCACATTTCATGTCACTTGAAATTTCATGCCT-GTATATGAG------  1200
L36156      AGGCACATTTCATGTCACTTGAAATCCTATGCCTTGTATATTAGAGGACG  1177
X90692      -----CATGTGGTGTA--TTATGTTTTTGTTATGTTCTTCAAGTTGATCA  1161
                  **.  ....   .*...     .*.  . . .. ....

L78163      ----------------------------------------  1200
U41657      ATAATTATTTGAATCTC-------AAAAAAAAAAAAAAAA  1031
X90693      AAAATCTTTTGGATTTC----ATTTGAAGTGTTTCT----  1200
X90694      ----------------------------------------  1200
L36156      TGT-TCTT--------C------TTGGTATTATACTA--T  1200
X90692      GGGA-CTGTAGAAGCTCCCTAATAATATTTGTGTCAAAGT  1200
```

FIGURE 3B

```
L78163    MGSMRLLVVALLCAFAMHAGFSVSY---AQLTPTFYRETCPNLFPIVFGV    47
U41657    --------------------------------------------------     0
X90693    MNSLRAVAIALCCIV--VVLGGLPFSSNAQLDPSFYRNTCPNVSSIVREV    48
X90694    MNSL---ATSMWCVVLLVVLGGLPFSSDAQLSPTFYSKTCPTVSSIVSNV    47
L36156    M----------WCVVLLVVLGGLPFSSDAQLSPTFYSKTCPTVSSIVSNV    40
X90692    MLGLSATA---FCCMVFVLIGGVPFS-NAQLDPSFYNSTCSNLDSIVRGV    46

L78163    IFDASFTDPRIGASLMRLHFHDCFVQGCDGSVLLNNTDTIESEQDALPNI    97
U41657    ------------------FHDCFVQGCDGSVLLNNTDTIESEQDALPNI    31
X90693    IRSVSKKDPRMLASLVRLHFHDCFVQGCDASVLLNKTDTVVSEQDAFPNR    98
X90694    LTNVSKTDPRMLASLVRLHFHDCFVLGCDASVLLNNTATIVSEQQAFPNN    97
L36156    LTNVSKTDPRMLASLVRLHFHDCFVLGCDASVLLNNTATIVSEQQAFPNN    90
X90692    LTNVSQSDPRMLGSLIRLHFHDCFVQGCDASILLNDTATIVSEQSAPPNN    96
                          ****** ***.*.***.*.*. *** *.**

L78163    NSIRGLDVVNDIKTAVENSCPDTVSCADILAIAAEIASVLGGGPGWPVPL   147
U41657    NSIRGLDVVNDIKTAVENSCPDTVSCADILAIAAEIASVAGRRSGWPVPL    81
X90693    NSLRGLDVVNQIKTAVEKACPNTVSCADILALSAELSSTLADGPDWKVPL   148
X90694    NSLRGLDVVNQIKLAVEVPCPNTVSCADILALAAQASSVLAQGPSWTVPL   147
L36156    NSLRGLDVVNQIKTAVESACPNTVSCADILALA-QASSVLAQGPSWTVPL   139
X90692    NSIRGLDVINQIKTAVENACPNTVSCADILALSAEISSDLANGPTWQVPL   146
          **.*****.*.** *** .**.*********.. . .*  .  ..* ***

L78163    GRRDSLTANRTLANQNLPAPFFNLTQLKASFAVQGLNTLDLVTLSGGHTF   197
U41657    GRRDSLTANRTLANQNLPAPFFNLTQLKASFAVQGLNTLDLVTLSGGHTS   131
X90693    GRRDGLTANQLLANQNLPAPFNTTDQLKAAFAAQGLDTTDLVALSGAHTF   198
X90694    GRRDGLTANRTLANQNLPAPFNSLDQLKAAFTAQGLNTTDLVALSGAHTF   197
L36156    GRRDGLTANRTLANQNLPAPFNSLDHLKLHLTAQGLITPVLVALSGAHTF   189
X90692    GRRDSLTANNSLAAQNLPAPTFNLTRLKSNFDNQNLSTTDLVALSGGHTI   196
          ****.****. **.******  . ..**  .. *.* *  **.***.**

L78163    GRARCSTFINRLYNFSNTGNPDPTLNTTYLEVLRARCPQNATGDNLTNLD   247
U41657    GRARCSTFINRLYNFSNTGLIH--LDTTYLEVLRARCPQNATGDNLTNLD   179
X90693    GRAHCSLFVSRLYNFSGTGSPDPTLNTTYLQQLRTICPNGGPGTNLTNFD   248
X90694    GRAHCAQFVSRLYNFSSTGSPDPTLNTTYLQQLRTICPNGGPGTNLTNFD   247
L36156    GRAHCAQFVSRLYNFSSTGSPDPTLNTTYLQQLRTICPNGGPGTNLTNFD   239
X90692    GRGQCRFFVDRLYNFSNTGNPDSTLNTTYLQTLQAICPNGGPGTNLTDLD   246
          **..*  *..******.**  .  *.****. *.. **....*.***..*

L78163    LSTPDQFDNRYYSNLLQLNGLLQSDQELFSTPGADTIPIVNSFSSNQNTF   297
U41657    LSTPDQFDNRYYSNLLQLNGLLQSDQERFSTPGADTIPLSIA-SANQNTF   228
X90693    PTTPDKFDKNYYSNLQVKKGLLQSDQELFSTSGSDTISIVNKFATDQKAF   298
X90694    PTTPDKFDKNYYSNLQVKKGLLQSDQELFSTSGADTISIVNKFSTDQNAF   297
L36156    PTTPDKFDKNYYSNLQVKKGLLQSDQELFSTSGADTISIVDKFSTDQNAF   289
X90692    PTTPDTFDSNYYSNLQVGKGLFQSDQELFSRNGSDTISIVNSFANNQTLF   296
           .*** **..*****   .**.***** **  *.***..    ...*. *

L78163    FSNFRVSMIKMGNIGVLTGDEGEIRLQCNFVN-----GDSFGLASVAS-K   341
U41657    FSNFRVSMIKMGNIGVLTGDEGEIRLQCNFVN-----GDSFGLASVAS-K   272
X90693    FESFRAAMIKMGNIGVLTGNQGEIRKQCNFVN---SKSAELGLINVAS-A   344
X90694    FESFKAAMIKMGNIGVLTGTKGEIRKQCNFVNFVNSNSAELDLATIASIV   347
L36156    FESFKAAMIKMGNIGVLTGTKGEIRKQCNFVN---SNSAELDLATIASIV   336
```

```
X90692      FENFVASMIKMGNIGVLTGSQGEIRTQCNAVN-----GNSSGLATVVT-K   340
            *..* ..************..**** *** **      ... .* ....

L78163      DAKQKLVAQSK   352
U41657      DAKQKLVAQSK   283
X90693      DSSEEGMVSSM   355
X90694      ESLEDGIASVI   358
L36156      ESLEDGIASVI   347
X90692      ESSEDGMASSF   351
            .. .. ..
```

FIGURE 4

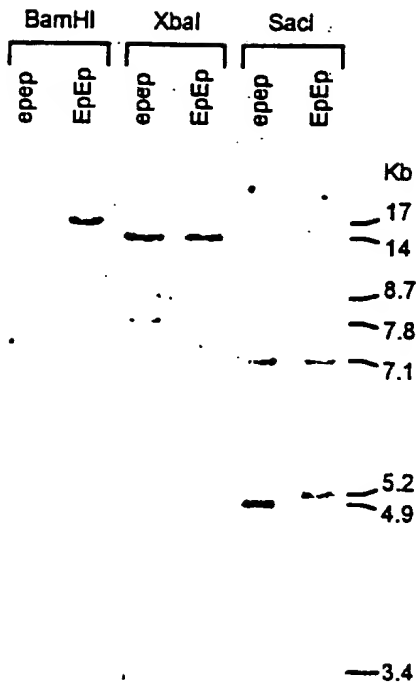

BamHI
XbaI
SacI
epep
EpEp
epep
EpEp
epep
EpEp
Kb
17
14
8.7
7.8
7.1
5.2
4.9
3.4

FIGURE 5

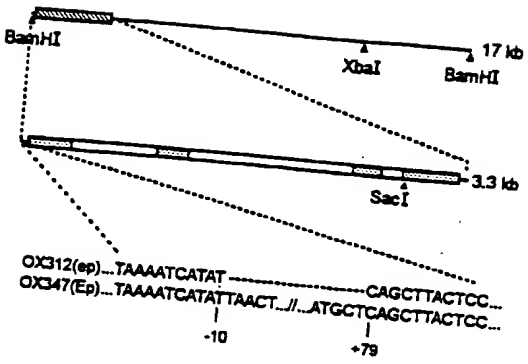

BamHI
XbaI
BamHI
17 kb
3.3 kb
SacI
OX312(ep)...TAAAATCATAT
OX347(Ep)...TAAAATCATATTAACT...//...ATGCTCAGCTTACTCC...
CAGCTTACTCC...
-10
+79

FIGURE 6

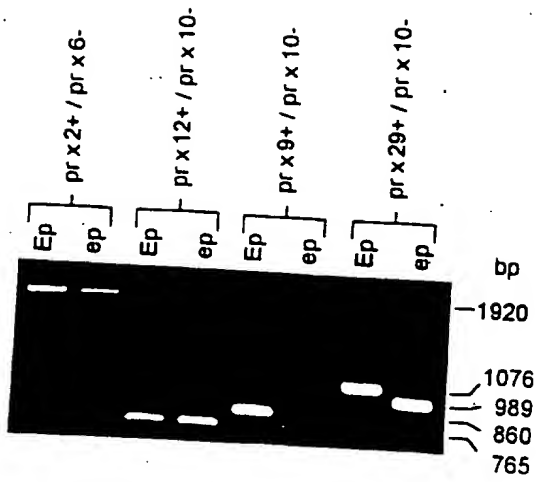

pr x 2+ / pr x 6-
pr x 12+ / pr x 10-
pr x 9+ / pr x 10-
pr x 29+ / pr x 10-
Ep
ep
bp
1920
1076
989
860
765

FIGURE 7

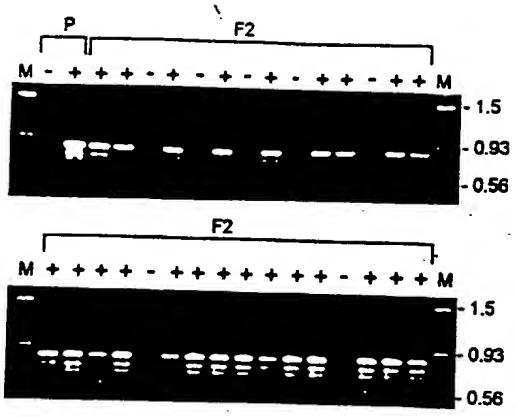


FIGURE 8

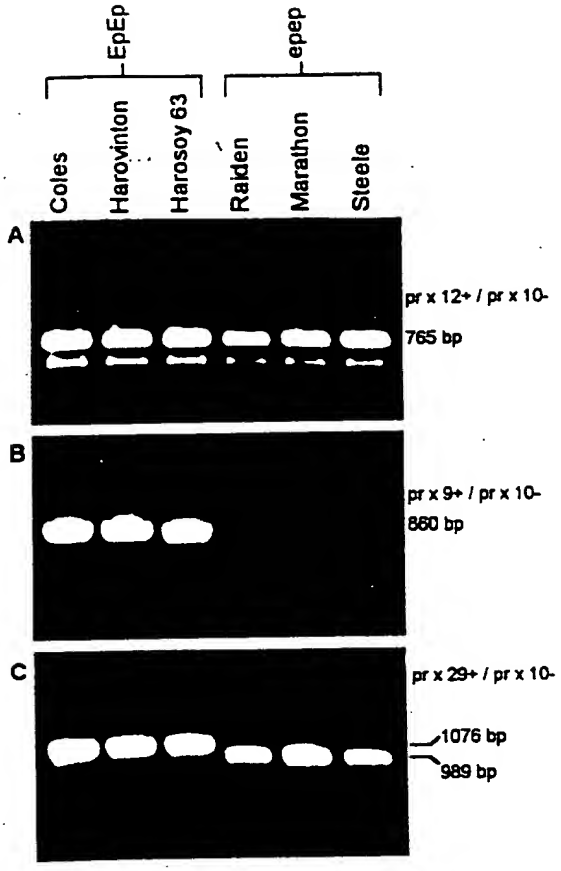
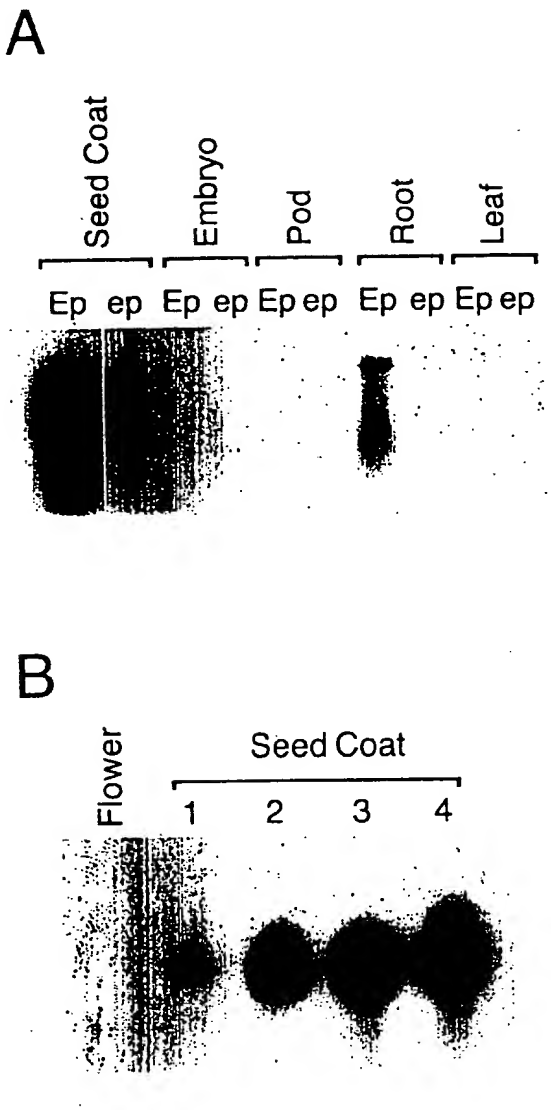

EpEp
epep
Coles
Harovinton
Harosoy 63
Raiden
Marathon
Steele
A
pr x 12+ / pr x 10-
765 bp
B
pr x 9+ / pr x 10-
860 bp
C
pr x 29+ / pr x 10-
1076 bp
989 bp

FIGURE 9


A
Seed Coat
Embryo
Pod
Root
Leaf
Ep ep Ep ep Ep ep Ep ep Ep ep
B
Flower
Seed Coat
1 2 3 4

OPIC
Office de la propriété intellectuelle du Canada

CIPO
Canadian Intellectual Property Office

Ottawa Hull K1A 0C9

(21) (A1) 2,149,000

(22) 1995/05/09

(43) 1996/11/10

(51) Int.Cl.[6] C12N 15/11; C12N 15/84; C12N 5/10; A01H 5/00

(19)(CA) **APPLICATION FOR CANADIAN PATENT** (12)

(54) Seed Coat-Specific Cryptic Promoter in Tobacco

(72) Miki, Brian - Canada ;
Fobert, Pierre - Canada ;
Iyer, V. N. - Canada ;

(71) Same as inventor

(57) 14 Claims

ABSTRACT

T-DNA tagging with a promoterless β-glucuronidase (GUS) gene generated a transgenic *Nicotiana tabacum* plant that expressed GUS activity only in developing seed coats. Cloning and deletion analysis of the GUS fusion revealed that the promoter responsible for seed coat specificity was located in the plant DNA proximal to the GUS gene. Analysis of the region demonstrated that the seed coat-specificity of GUS expression in this transgenic plant resulted from T-DNA insertion next to a cryptic promoter. This promotor is useful in controlling the expression of genes to the developing seed coat in plant seeds.

A SEED COAT-SPECIFIC CRYPTIC PROMOTER IN TOBACCO

Field of Invention

This invention relates to a cryptic promoter identified from *Nicotiana tabacum* (tobacco). Specifically this invention relates to a seed coat-specific cryptic promoter isolated from tobacco.

Background and Prior Art

Bacteria from the genus *Agrobacterium* have the ability to transfer specific segments of DNA (T-DNA) to plant cells, where they stably integrate into the nuclear chromosomes. Analyses of plants harbouring the T-DNA have revealed that this genetic element may be integrated at numerous locations, and can occasionally be found within genes. One strategy which may be exploited to identify integration events within genes is to transform plant cells with specially designed T-DNA vectors which contain a reporter gene, devoid of *cis*-acting transcriptional and translational expression signals (i.e. promoterless), located at the end of the T-DNA. Upon integration, the initiation codon of the promoterless gene (reporter gene) will be juxtaposed to plant sequences. The consequence of T-DNA insertion adjacent to, and downstream of, gene promoter elements may be the activation of reporter gene expression. The resulting hybrid genes, referred to as T-DNA-mediated gene fusions, consist of unknown and thus un-characterized plant promoters residing at their natural location within the chromosome, and the coding sequence of a marker gene located on the inserted T-DNA (Fobert *et al.*, 1991, Plant Mol. Biol. 17, 837-851).

It has generally been assumed that activation of promoterless or enhancerless marker genes result from T-DNA insertions within or immediately adjacent to genes. The recent isolation of several T-DNA insertional mutants (Koncz *et al.*, 1992, *Plant Mol. Biol.* 20, 963-976; reviewed in Feldmann, 1991, *Plant J.* 1, 71-82; Van Lijsebettens *et al.*,

1991, *Plant Sci.* 80, 27-37; Walden *et al.*, 1991, *Plant J.* 1: 281-288; Yanofsky *et al.*, 1990, *Nature* 346, 35-39), shows that this is the case for at least some insertions. However, other possibilities exist. One of these is that integration of the T-DNA activates silent regulatory sequences that are not associated with genes. Lindsey *et al.* (1993, *Transgenic Res.* 2, 33-47) referred to such sequences as "pseudo-promoters" and suggested that they may be responsible for activating marker genes in some transgenic lines.

Inactive regulatory sequences that are buried in the genome but with the capability of being functional when positioned adjacent to genes have been described in a variety of organisms, where they have been called "cryptic promoters" (Al-Shawi *et al.*, 1991, *Mol. Cell. Biol.* 11, 4207-4216; Fourel *et al.*, 1992, *Mol. Cell. Biol.* 12, 5336-5344; Irniger *et al.*, 1992, *Nucleic Acids Res.* 20, 4733-4739; Takahashi *et al.*, 1991, *Jpn J. Cancer Res.* 82, 1239-1244). Cryptic promoters can be found in the introns of genes, such as those encoding for yeast actin (Irniger *et al.*, 1992, *Nucleic Acids Res.* 20, 4733-4739), and a mammalian melanoma-associated antigen (Takahashi *et al.*, 1991, *Jpn J. Cancer Res.* 82, 1239-1244). It has been suggested that the cryptic promoter of the yeast actin gene may be a relict of a promoter that was at one time active but lost function once the coding region was assimilated into the exon-intron structure of the present-day gene (Irniger *et al.*, 1992, *Nucleic Acids Res.* 20, 4733-4739). A cryptic promoter has also been found in an untranslated region of the second exon of the woodchuck N-myc proto-oncogene (Fourel *et al.*, 1992, *Mol. Cell. Biol.* 12, 5336-5344). This cryptic promoter is responsible for activation of a N-myc2, a functional processed gene which arose from retropositon of N-*myc* transcript (Fourel *et al.*, 1992, *Mol. Cell. Biol.* 12, 5336-5344). These types of regulatory sequences have not yet been isolated from plants.

This patent application describes, as an example, ne transgenic plant, T218, generated by tagging with a promoterless GUS (β-glucuronidase) T-DNA vector. This plant is of particular interest in that GUS expression was spatially and developmentally regulated in seed coats and a promoter specific to this tissue has not been previously isolated. Cloning of the insertion site uncovered a cryptic promoter within a region of the tobacco genome not conserved among related species. This seed coat-specific promoter can be useful for controlling gene expression of selected genes to a specific stage of development.

Summary of Invention

The present invention is directed to a cryptic promoter identified from *Nicotiana tabacum* (tobacco). Specifically this invention relates to a seed coat-specific cryptic promoter isolated from tobacco.

The transgenic tobacco plant, T218, contained a 4.7 kb *Eco*RI fragment containing the 2.2 kb promoterless GUS-*nos* gene and 2.5 kb of 5' flanking tobacco DNA. Deletion of the region approximately between 2.5 and 1.0 kb of the 5' flanking region did not alter GUS expression, as compared to the entire 4.7 kb GUS fusion. A further deletion to 0.5 kb of the 5' flanking site resulted in complete lose of GUS activity. Thus the region between 1.0 and 0.5 of the 5' flanking region of the tobacco DNA contains the elements essential to gene activation. This region is contained within a *Xba*I - *Sna*BI restriction site fragment of the flanking tobacco DNA.

Thus according to the present invention there is provided a seed coat-specific cryptic promoter in tobacco contained within a DNA sequence, or analogue thereof, as shown in Figure 6.

Further according t the present invention, there is provided a DNA sequence, or analogue thereof, as shown in Figure 6.

This invention also relates to a cloning vector containing a seed coat-specific cryptic promoter from tobacco, which is contained within a DNA sequence, or analogue thereof, as shown in Figure 6 and a gene encoding a protein.

This invention also includes a plant cell which has been transformed with a cloning vector as described above

This invention further relates to a transgenic plant containing a seed-coat specific promoter, operatively linked to a gene encoding a protein.

Brief Description of the Drawings

Figure 1 depicts the fluorogenic analyses of GUS expression in the plant T218. Each bar represents the average ± one standard deviation of three samples. Nine different tissues were analyzed: leaf (L), stem (S), root (R), anther (A), petal (P), ovary (O), sepal (Se), seeds 10 days post anthesis (S1) and seeds 20 days post-anthesis (S2). For all measurements of GUS activity, the fraction attributed to intrinsic fluorescence, as determined by analysis of untransformed tissues, is shaded black on the graph. Absence of a black area at the bottom of a histogram indicates that the relative contribution of the background fluorescence is too small to be apparent.

Figure 2 shows the cloning of the GUS fusion in plant T218 (pT218) and construction of transformation vectors. Plant DNA is indicated by the solid line and the promoterless GUS-*nos* gene is indicated by the open box. The transcriptional start site and presumptive TATA box

are located by the closed and open arrow heads respectively. DNA probes #1, 2, 3 and RNA probe #4 are shown. The *Eco*RI fragment in pT218 was subcloned in the pBIN19 polylinker to create pT218-1. Fragments truncated at the *Xba*I *Sna*BI and *Xba*I sites were also subcloned to create pT218-2, pT218-3 and pT218-4. Abbreviations for the endonuclease restriction sites are as follows: *Eco*RI (E), *Hin*dIII (H), *Xba*I (X), *Sna*BI (N), *Sma*I (M), *Sst*I (S).

Figure 3 shows the expression pattern of promoter fusions during seed development. GUS activity in developing seeds (4-20 days postanthesis (dpa)) of (Fig. 3a) plant T218 (●-●) and (Fig. 3b) plants transformed with vectors pT218-1 (○-○), pT218-2 (□-□), pT218-3 (▽-▽) and pT218-4 (△-△) which are illustrated in Figure 2. The 2 day delay in the peak of GUS activity during seed development, seen with the pT218-2 transformant, likely reflects greenhouse variation conditions.

Figure 4 shows GUS activity in 12 dpa seeds of independent transformants produced with vectors pT218-1 (○), pT218-2 (□), pT218-3 (▽) and pT218-4 (△). The solid markers indicate the plants shown in Figure 3 (b) and the arrows indicate the average values for plants transformed with pT218-1 or pT218-2.

Figure 5 shows the mapping of the T218 GUS fusion termini and expression of the region surrounding the insertion site in untransformed plants.

(Fig. 5a) Mapping of the GUS mRNA termini in plant T218. The antisense RNA probe from subclone #4 (Figure 2) was used for hybridization with total RNA of tissues from untransformed plants (10 μg) and from plant T218 (30 μg). Arrowheads indicate the anticipated position of protected fragments if

transcripts were initiated at the same sites as the T218 GUS fusion.

(Fig. 5b) RNase protection assay using the antisense (relative to the orientation of the GUS coding region) RNA probe from subclone e (Figure 7) against 30 μg total RNA of tissues from untransformed plants.

P, untreated RNA probe; -, control assay using the probe and tRNA only; L, leaves from untransformed plants; 8, 10, 12, seeds from untransformed plants at 8, 10, and 12 dpa, respectively; T10, seeds of plant T218 at 10 dpa; +, control hybridization against unlabeled *in vitro*-synthesized sense RNA from subclone c (panel a) or subclone e (panel b). The two hybridizing bands near the top of the gel are end-labeled DNA fragment of 3313 and 1049 bp, included in all assays to monitor losses during processing. Molecular weight markers are in number of bases.

Figure 6 provides the nucleotide sequence of pT218 (top line) and pIS-1 (bottom line). Sequence identity is indicated by dashed lines. The T-DNA insertion site is indicated by a vertical line after bp 993. This site on pT218 is immediately followed by a 12 bp filler DNA, which is followed by the T-DNA. The first nine amino acids of the GUS gene and the GUS initiation codon (*) are shown. The major and minor transcriptional start site is indicated by a large and small arrow, respectively. The presumptive TATA box is identified and is in boldface. Additional putative TATA and CAAT boxes are marked with boxes. The location of direct (1-5) and indirect (6-8) repeats are indicated by arrows.

Figure 7 shows the base composition of region surrounding the T218 insertion site cloned from untransformed plants. The site of T-DNA insertion in plant T218 is indicated by the vertical arrow. The position of the 2 genomic clones pIS-1 and pIS-2, and of the various RNA probes (a-e) used in RNase protection assays are indicated beneath the graph.

Figure 8 shows the Southern blot analyses of the insertion site in *Nicotiana* species. DNA from *N. tomentosiformis* (N tom), *N. sylvestris* (N syl), and *N. tabacum* (N tab) were digested with *Hin*dIII (H), *Xba*I (X) and *Eco*RI (E) and hybridized using probe #2 (Figure 2). Lambda *Hin*dIII markers (kb) are indicated.

Figure 9 shows the AT content of 5' non-coding regions of plant genes. A program was written in PASCAL to scan GenBank release 75.0 and to calculate the AT contents of the 5' non-coding (solid bars) and the coding regions (hatched bars) of all plant genes identified as "Magnoliophyta" (flowering plants). The region -200 to -1 and +1 to +200 were compared. Shorter sequences were also accepted if they were at least 190 bp long. The horizontal axis shows the ratio of the AT content (%). The vertical axis shows the number of the sequences having the specified AT content ratios.

Detailed Description of the Preferred Embodiments

T-DNA tagging with a promoterless β-glucuronidase (GUS) gene generated a transgenic *Nicotiana tabacum* plant that expressed GUS activity only in developing seed coats. Cloning and deletion analysis of the GUS fusion revealed that the promoter responsible for seed coat specificity was located in the plant DNA proximal to the GUS gene. Deletion analyses localized the cryptic promoter to an approximately 0.5 kb region between a *Xba*I and a *Sna*BI restriction endonuclease site of the 5' flanking tobacco DNA. This region spans from nucleotide 1 to nucleotide 467 as shown in Figure 6.

Thus, the present invention includes a DNA sequence comprising the seed coat-specific cryptic promoter from tobacco and analogues, thereof. Analogues of the cryptic promoter include any substitution,

deletion, or additions of the region, provided that said analogues maintain the seed coat- specific expression activity.

The term cryptic promoter means a promoter that is not associated with a gene and thus does not control expression in its native location. These inactive regulatory sequences are buried in the genome but are capable of being functional when positioned adjacent to a gene.

The DNA sequence of the present invention thus includes the DNA sequence of as shown in Figure 6, the promoter region within the sequence as shown in Figure 6 (for example from nucleotide 1 to 476), and analogues thereof. Analogues include those DNA sequences which hybridize under stringent hybridization conditions (see Maniatis *et al.*, in Molecular Cloning (A Laboratory Manual), Cold Spring Harbor Laboratory, 1982, p. 387-389) to the DNA sequence as shown in Figure 6, provided that said sequences maintain the seed coat-specific promoter activity. An example of one such stringent hybridization conditions may be hybridization at 4XSSC at 65°C, followed by washing in 0.1XSSC at 65°C for an hour. Alternatively an exemplary stringent hybridization condition could be in 50% formamide, 4XSSC at 42°C. Analogues also include those DNA sequences which hybridize to the sequence as shown in Figure 6 under relaxed hybridization conditions, provided that said sequences maintain the seed coat-specific promoter activity. Examples of such non-hybridization conditions includes hybridization at 4XSSC at 50°C or with 30-40% formamide at 42°C.

There are several lines of evidence that suggest that the seed coat-specific expression of GUS activity in the plant T218 is regulated by a cryptic promoter. The region surrounding the promoter and transcriptional start site for the GUS gene are not transcribed in untransformed plants. Transcription was only observed in plant T218 when T-DNA was inserted in *cis*. DNA sequence analysis did not uncover

a long open reading frame within the 3.3 kb region cloned. Moreover, the region is very AT rich and predicted to be noncoding (data not shown) by the Fickett algorithm (Fickett, 1982, *Nucleic Acids Res.* 10, 5303-5318) as implemented in DNASIS 7.0 (Hitachi). Southern blots revealed that the insertion site is within the *N. tomentosiformis* genome and is not conserved among related species as would be expected for a region with an important gene.

As this is the first report of a cryptic promoter in plants, it is impossible to estimate the degree to which cryptic promoters may contribute to the high frequencies of promoterless marker gene activation in plants. It is interesting to note that transcriptional GUS fusions in *Arabidopsis* occur at much greater frequencies (54%) than translational fusions (1.6%, Kertbundit *et al.*, 1991, *Proc. Natl. Acad. Sci. USA* 88, 5212-5216). The possibility that cryptic promoters may account for some fusions was recognized by Lindsey *et al.* (1993, *Transgenic Res.* 2, 33-47).

The results disclosed herewith confirms others (Gheysen *et al.*, 1987, *Proc. Natl. Acad. Sci. USA* 84, 6169-6173 and 1991, *Genes Dev.* 5, 287-297) that T-DNA may insert into A-T rich regions as do plant transposable elements (Capel *et al.*, 1993, *Nucleic Acids Res.* 21, 2369-2373). We illustrate that promoters of plant genes are also A-T rich raising speculation that gene insertions into these regions could facilitate the rapid acquisition of new regulatory elements during gene evolution.

The insertion of functional genes into the nuclear genome and acquisition of new regulatory sequences has already played a major role in the diversification of certain genes and the endosymbiosis of organelles. In plants, most organellar proteins are nuclear encoded due to the ongoing transfer of their genes into the nucleus (Palmer, 1991, In Bogorad L and Vasil IK (eds) The Molecular Biology of Plastids, Academic Press, San Diego, pp 5-53). Recently, it has been shown that the *cox 2* gene of

cowpea (Nugent and Palmer, 1991, *Cell* 66, 473-481) and soybean (Covell and Gray, 1992, *EMBO J.* 11, 3815-3820) were transferred from mitochondria to nucleus without promoters by RNA intermediates. The results disclosed herewith, with T-DNA-mediated gene fusions reveal the facility with which promoters can be acquired by incoming genes. The presence of cryptic promoters and diverse regulatory elements in the intergenic regions may insure that genes rapidly achieve the features needed to meet the demands of complex multicellular organisms.

The cryptic promoter of the present invention can also be used to control to the expression of any given gene spatially and developmentally to developing seed coats. Some examples of such uses, which are not to be considered limiting, include:

1. Modification of storage reserves in seed coats, such as starch by the expression of yeast invertase to mobilize the starch or expression of the antisense transcript of ADP-glucose pyrophosphorylase to inhibit starch biosynthesis.

2. Modification of seed color contributed by condensed tannins in the seed coats by expression of antisense transcripts of the phenylalanine ammonia lyase or chalcone synthase genes.

3. Modification of fibre content in seed-derived meal by expression of antisense transcripts of the caffeic acid-o-methyl transferase or cinnamoyl alcohol dehydrogenase genes.

4. Inhibition of seed coat maturation by expression of ribonuclease genes to allow for increased seed size, and to reduce the relative biomass of seed coats, and to aid in dehulling of seeds.

5. Expression f genes in seed coats coding f r insecticidal proteins such as α-amylase inhibitor or protease inhibitor.

6. Partitioning of seed metabolites such as glucosinolates into seed coats for nematode resistant.

Thus this invention is directed to such promoter and gene combinations. Further this invention is directed to such promoter and gene combinations in a cloning vector, wherein the gene is under the control of the promoter and is capable of being expressed in a plant cell transformed with the vector. This invention further relates to transformed plant cells and transgenic plants regenerated from such plant cells. The promoter and promoter gene combination of the present invention can be used to transform any plant cell for the production of any transgenic plant. The present invention is not limited to any plant species.

While this invention is described in detail with particular reference to preferred embodiments thereof, said embodiments are offered to illustrate but not limit the invention.

EXAMPLES

Characterization of a Seed Coat-Specific GUS Fusion

Transfer of binary constructs to *Agrobacterium* and leaf disc transformation of *Nicotiana tabacum* SR1 were performed as described by Fobert *et al.* (1991, *Plant Mol. Biol.* 17, 837-851). Plant tissue was maintained on 100 μg/ml kanamycin sulfate (Sigma) throughout *in vitro* culture.

Nine-hundred and forty transgenic plants were produced. Several hundred independent transformants were screened for GUS activity in

developing seeds using the fluorogenic assay. One of these, T218, was chosen for detailed study because of its unique pattern of GUS expression.

Fluorogenic and histological GUS assays were performed according to Jefferson (*Plant Mol. Biol. Rep.*, 1987, **5**, 387-405), as modified by Fobert *et al.* (*Plant Mol. Biol.*, 1991, **17**, 837-851). For initial screening, leaves were harvested from *in vitro* grown plantlets. Later flowers corresponding to developmental stages 4 and 5 of Koltunow *et al.* (*Plant Cell*, 1990, **2**, 1201-1224) and beige seeds, approximately 12-16 dpa (Chen *et al.*, 1988, *EMBO J.* **7**, 297-302), were collected from plants grown in the greenhouse. For detailed, quantitative analysis of GUS activity, leaf, stem and root tissues were collected from kanamycin resistant F1 progeny of the different transgenic lines grown *in vitro*. Floral tissues were harvested at developmental stages 8-10 (Koltunow *et al.*, 1990, *Plant Cell* **2**, 1201-1224) from the original transgenic plants. Flowers of these plants were also tagged and developing seeds were collected from capsules at 10 and 20 dpa. In all cases, tissue was weighed, immediately frozen in liquid nitrogen, and stored at -80°C.

Tissues analyzed by histological assay were at the same developmental stages as those listed above. Different hand-cut sections were analyzed for each organ. For each plant, histological assays were performed on at least two different occasions to ensure reproducibility. Except for floral organs, all tissues were assayed in phosphate buffer according to Jefferson (1987, *Plant Mol. Biol. Rep.* **5**, 387-405), with 1 mM X-Gluc (Sigma) as substrate. Flowers were assayed in the same buffer containing 20% (v/v) methanol (Kosugi *et al.*, 1990, *Plant Sci.* **70**, 133-140).

Tissue-specific patterns of GUS expression were only found in seeds. For instance, GUS activity in plant T218 (Figure 1) was localized in seeds from 9 to 17 days postanthesis (dpa). GUS activity was not detected

in seeds at ther stages of development or in any ther tissue analyzed which includ d leaf, stem, root, anther, ovary, petal and sepal (Figure 1). Histological staining with X-Gluc revealed that GUS expression in seeds at 14 dpa was localized in seed coats but was absent from the embryo, endosperm, vegetative organs and floral organs (results not shown).

The seed coat-specificity of GUS expression was confirmed with the more sensitive fluorogenic assay of seeds derived from reciprocal crosses with untransformed plants. The seed coat differentiates from maternal tissues called the integuments which do not participate in double fertilization (Esau, 1977, *Anatomy of Seed Plants.* New York: John Wiley and Sons). If GUS activity is strictly regulated, it must originate from GUS fusions transmitted to seeds maternally and not by pollen. As shown in Table 1, this is indeed the case. As a control, GUS fusions expressed in embryo and endosperm, which are the products of double fertilization, should be transmitted through both gametes. This is illustrated in Table 1 for GUS expression driven by the napin promoter (BngNAPI, Baszczynki and Fallis, 1990, *Plant Mol. Biol.* 14, 633-635) which is active in both embryo and endosperm (data not shown).

Table 1. GUS activity in seeds at 14 days post anthesis.

Cross ♀	Cross ♂	GUS Activity nmole MU/min/mg Protein
T218	T218	1.09 ± 0.39
T218	WT[a]	3.02 ± 0.19
WT	T218	0.04 ± 0.005
WT	WT	0.04 ± 0.005
NAP-5[b]	NAP-5	14.6 ± 7.9
NAP-5	WT	3.42 ± 1.60
WT	NAP-5	2.91 ± 1.97

[a] WT, untransformed plants

[b] Transgenic tobacco plants with the GUS gene fused to the napin, BngNAP1, promoter (Baszczynski and Fallis, 1990, *Plant Mol. Biol.* 14, 633-635).

Cloning and Analysis of the Seed Coat-Specific GUS Fusion

Genomic DNA was isolated from freeze-dried leaves using the protocol of Sanders *et al.* (1987, *Nucleic Acid Res.* 15, 1543-1558). Ten micrograms of T218 DNA was digested for several hours with *Eco*RI using the appropriate manufacturer-supplied buffer supplemented with 2.5 mM spermidine. After electrophoresis through a 0.8% TAE agarose gel, the DNA size fraction around 4-6 kb was isolated, purified using the GeneClean kit (BIO 101 Inc., LaJolla, CA), ligated to phosphatase-treated *Eco*RI-digested Lambda GEM-2 arms (Promega) and packaged *in vitro* as suggested by the supplier. Approximately 125,000 plaques were transferred to nylon filters (Nytran, Schleicher and Schuell) and screened by plaque hybridization (Rutledge *et al.*, 1991, *Mol. Gen. Genet.* 229, 31-40), using the 3' (termination signal) of the *nos* gene as probe (probe #1, Figure 2). This sequence, contained in a 260 bp *Sst*I/*Eco*RI restriction fragment from pPRF-101 (Fobert *et al.*, 1991, *Plant Mol. Biol.* 17, 837-851), was labeled with [α-^{32}P]-dCTP (NEN) using random priming (Stratagene).

After plaque purification, phage DNA was isolated (Sambrook *et al.*, 1989, A Laboratory Manual. New York: Cold Spring Harbor Laboratory Press), mapped and subcloned into pGEM-4Z (Promega). The *Eco*RI fragment and deletions shown in Figure 2 were inserted into pBIN19 (Bevan, 1984, *Nucl. Acid Res.* 12, 8711-8721). Restriction mapping was used to determine the orientation of the fusion in pBIN19 and to confirm plasmid integrity. Plants were transformed with a derivative which contained the 5' end of the GUS gene distal to the left border repeat. This orientation is the same as that of the GUS gene in the binary vector pBI101 (Jefferson, 1987, *Plant Mol. Biol. Rep.* 5, 387-405).

The GUS fusion in plant T218 was isolated as a 4.7 kb *Eco*RI fragment containing the 2.2kb promoterless GUS-*nos* gene at the T-DNA border of pPRF120 and 2.5 kb of 5' flanking tobacco DNA (pT218, Figure 2), using the *nos* 3' fragment as probe (probe #1, Figure 2). To confirm the ability of the flanking DNA to activate the GUS coding region, the entire 4.7 kb fragment was inserted into the binary transformation vector pBIN19 (Bevan, 1984, *Nucl. Acid Res.* 12, 8711-8721), as shown in Figure 2. Several transgenic plants were produced by *Agrobacterium*-mediated transformation of leaf discs. Southern blots indicated that each plant contained 1-4 T-DNA insertions at unique sites. The spatial patterns of GUS activity were identical to that of plant T218. Histologically, GUS staining was restricted to the seed coats of 14 dpa seeds and was absent in embryos and 20 dpa seeds (results not shown). Fluorogenic assays of GUS activity in developing seeds showed that expression was restricted to seeds between 10 and 17 dpa, reaching a maximum at 12 dpa (Figure 3 (a) and 3 (b)). The 4.7 kb fragment therefore contained all of the elements required for the tissue-specific and developmental regulation of GUS expression.

To locate regions within the flanking plant DNA responsible for seed coat-specificity, truncated derivatives of the GUS fusion were generated (Figure 2) and introduced int tobacco plants. Deletion of the

region approximately between 2.5 and 1.0 kb, 5' of the insertion site (pT218-2, Figure 2) did not alter expression compared with the entire 4.7 kb GUS fusion (Figures 3b and 4). Further deletion of the DNA, to the *Sna*BI restriction site approximately 0.5 kb, 5' of the insertion site (pT218-3, Figure 2), resulted in the complete loss of GUS activity in developing seeds (Figures 3b and 4). This suggests that the region approximately between 1.0 and 0.5 kb, 5' of the insertion site contains elements essential to gene activation. GUS activity in seeds remained absent with more extensive deletion of plant DNA (pT218-4, Figures 2, 3b and 4) and was not found in other organs including leaf, stem, root, anther, petal, ovary or sepal from plants transformed with any of the vectors (data not shown).

The transcriptional start site for the GUS gene in plant T218 was determined by RNase protection assays with RNA probe #4 (Figure 2) which spans the T-DNA/plant DNA junction. For RNase protection assays, various restriction fragments from pIS-1, pIS-2 and pT218 were subcloned into the transcription vector pGEM-4Z as shown in Figures 7 and 2, respectively. A 440bp *Hin*dIII fragment of the tobacco acetohydroxyacid synthase *SURA* gene was used to detect *SURA* and *SURB* mRNA. DNA templates were linearized and transcribed *in vitro* with either T7 or SP6 polymerases to generate strand-specific RNA probes using the Promega transcription kit and [α-^{32}P]CTP as labeled nucleotide. RNA probes were further processed as described in Ouellet *et al.* (1992, *Plant J.* 2, 321-330). RNase protection assays were performed as described in Ouellet *et al.*, (1992, *Plant J.* 2, 321-330), using 10-30 μg of total RNA per assay. Probe digestion was done at 30°C for 15 min using 30 μg ml^{-1} RNase A (Boehringer Mannheim) and 100 units ml^{-1} RNase T1 (Boehringer Mannheim). Figure 5 shows that two termini were mapped in the plant DNA. The major 5' terminus is situated at an adenine residue, 122 bp upstream of the T-DNA insertion site (Figure 6). The sequence at this transcriptional start site is similar to the consensus sequence for plant genes (C/TTC↓ATCA; Joshi, 1987 *Nucleic Acids Res.* 15, 6643-6653). A

TATA box consensus sequence is present 37 bp upstream f this start site (Figure 6). The second, minor terminus mapped 254 bp from the insertion site in an area where no obvious consensus motifs could be identified (Figure 6).

The tobacco DNA upstream of the insertion site is very AT-rich (>75%, see Figure 7). A search for promoter-like motifs and scaffold attachment regions (SAR), which are often associated with promoters (Breyne *et al.*, 1992, *Plant Cell* 4, 463-471; Gasser and Laemmli, 1986, *Cell* 46, 521-530), identified several putative regulatory elements in the first 1.0 kb of tobacco DNA flanking the promoterless GUS gene (data not shown). However, the functional significance of these sequences remains to be determined.

Cloning and Analysis of the Insertion Site from Untransformed Plants

A lambda DASH genomic library was prepared from DNA of untransformed *N. tabacum* SR1 plants by Stratagene for cloning of the insertion site corresponding to the gene fusion in plant T218. The screening of 500,000 plaques with probe #2 (Figure 2) yielded a single lambda clone. The *Eco*RI and *Xba*I fragments were subcloned in pGEM-4Z to generate pIS-1 and pIS-2. Figure 7 shows these two overlapping subclones, pIS-1 (3.0 kb) and pIS-2 (1.1 kb), which contain tobacco DNA spanning the insertion site (marked with a vertical arrow). DNA sequence analysis (using dideoxy nucleotides in both directions) revealed that the clones, pT218 and pIS-1, were identical over a length of more than 2.5 kb, from the insertion site to their 5' ends, except for a 12 bp filler DNA insert of unknown origin at the T-DNA border (Figure 6 and data not shown). The presence of filler DNA is a common feature of T-DNA/plant DNA junctions (Gheysen *et al.*, 1991, *Gene* 94, 155-163). Gross rearrangements that sometimes accompany T-DNA insertions (Gheysen *et al.*, 1990, *Gene* 94, 155-163; and 1991, *Genes Dev.* 5, 287-297) were not

found (Figure 6) and therefore could not account for the promoter activity associated with this region. The region f pIS-1 and pIS-2, 3' of the insertion site is also very AT-rich (Figure 7).

To determine whether there was a gene associated with the pT218 promoter, more than 3.3 kb of sequence contained with pIS-1 and pIS-2 was analyzed for the presence of long open reading frames (ORFs). However, none were detected in this region (data not shown). To determine whether the region surrounding the insertion site was transcribed in untransformed plants, Northern blots were performed with RNA from leaf, stem, root, flower and seeds at 4, 8, 12, 14, 16, 20 and 24 dpa. Total RNA from leaves was isolated as described in Ouellet *et al.*, (1992, *Plant J.* 2, 321-330). To isolate total RNA from developing seeds, 0.5 g of frozen tissue was pulverized by grinding with dry ice using a mortar and pestle. The powder was homogenized in a 50 ml conical tube containing 5 ml of buffer (1 M Tris HCl, pH 9.0, 1% SDS) using a Polytron homogenizer. After two extractions with equal volumes of phenol:chloroform:isoamyl alcohol (25:24:1), nucleic acids were collected by ethanol precipitation and resuspended in water. The RNA was precipitated overnight in 2M LiCl at 0°C, collected by centrifugation, washed in 70% ethanol and resuspended in water. Northern blot hybridization was performed as described in Gottlob-McHugh *et al.* (1992, *Plant Physiol.* 100, 820-825). Probe #3 (Figure 2) which spans the entire region of pT218 5' of the insertion did not detect hybridizing RNA bands (data not shown). To extend the sensitivity of RNA detection and to include the region 3' of the insertion site within the analysis, RNase protection assays were performed with 10 different RNA probes that spanned both strands of pIS-1 and pIS-2 (Figure 7). Even after lengthy exposures, protected fragments could not be detected with RNA from 8, 10, 12 dpa seeds or leaves of untransformed plants (see Figure 5 for examples with two of the probes tested). The specific conditions used allowed the resolution f protected RNA fragments as small as 10 bases

(data not shown). Failure to detect protected fragments was not due to problems f RNA quality, as control experiments using the same samples detected acetohydroxyacid synthase (*AHAS*) *SURA* and *SURB* mRNA which are expressed at relatively low abundance (data not shown). Conditions used in the present work were estimated to be sensitive enough to detect low-abundance messages representing 0.001-0.01% of total mRNA levels (Ouellet *et al.*, 1992, *Plant J.* 2, 321-330). Therefore, the region flanking the site of T-DNA insertion does not appear to be transcribed in untransformed plants.

Genomic Origins of the Insertion Site

Southern blots were performed to determine if the insertion site is conserved among *Nicotiana* species. Genomic DNA (5 μg) was isolated, digested and separated by agarose gel electrophoresis as described above. After capillary transfer on to nylon filters, DNA was hybridized, and probes were labeled, essentially as described in Rutledge *et al.* (1991, *Mol. Gen. Genet.* 229, 31-40). High-stringency washes were in 0.2 x SSC at 65°C while low-stringency washes were in 2 x SSC at room temperature. In Figure 8, DNA of the allotetraploid species *N. tabacum* and the presumptive progenitor diploid species *N. tomentosiformis* and *N. sylvestris* (Okamuro and Goldberg, 1985, *Mol. Gen. Genet.*, 198, 290-298) were hybridized with probe #2 (Figure 2). Single hybridizing fragments of identical size were detected in *N. tabacum* and *N. tomentosiformis* DNA digested with *Hin*dIII, *Xba*I and *Eco*RI, but not in *N. sylvestris*. Hybridizations with pIS-2 (Figure 8) which spans the same region but includes DNA 3' of the insertion site yielded the same results. They did not reveal hybridizing bands, even under conditions of reduced stringency, in additional *Nicotiana* species including *N. rustica*, *N. glutinosa*, *N. megalosiphon* and *N. debneyi* (data not shown). Probe #3 (Figure 2) revealed the presence of moderately repetitive DNA specific to the *N. tomentosiformis* genome (data not shown). These results suggest that the

region flanking the insertion site is unique to the *N. tomentosiformis* genome and is not conserved among related species as might be expected for regions that encode essential genes.

All scientific publications and patent documents are incorporated herein by reference.

The present invention has been described with regard to preferred embodiments. However, it will be obvious to persons skilled in the art that a number of variations and modifications can be made without departing form the scope of the invention as described in the following claims.

THE EMBODIMENTS OF THE INVENTION IN WHICH AN EXCLUSIVE PROPERTY OR PRIVILEGE IS CLAIMED ARE DEFINED AS FOLLOWS:

1. A seed coat-specific cryptic promoter from tobacco.

2. The promoter of claim 1, contained within a DNA sequence, or analogue thereof, as shown in Figure 6.

3. The promoter of claim 2, contained within a DNA sequence, or analogue thereof, from nucleotide 1 to nucleotide 467 as shown in Figure 6.

4. A DNA sequence, or analogue thereof, as shown in Figure 6, wherein said DNA sequence, or analog thereof, codes for a seed coat-specific promoter.

5. The sequence of claim 4, or analogue thereof, from nucleotide 1 to nucleotide 467 as shown in Figure 6.

6. A cloning vector which comprises a gene encoding a protein and a seed coat-specific cryptic promoter from tobacco, wherein the gene is under the control of the promoter and is capable of being expressed in a plant cell transformed with the vector.

7. The vector of claim 6, wherein the seed coat-specific promoter is contained within a DNA sequence, or analogue thereof, as shown in Figure 6.

8. The vector of claim 7, wherein the seed coat-specific promoter is contained within a DNA sequence, or analogue thereof, from nucleotide 1 to nucleotide 467 as shown in Figure 6.

9. A plant cell which has been transformed with a vector as claimed in claim 6.

10. A plant cell which has been transformed with a vector as claimed in claim 7.

11. A plant cell which has been transformed with a vector as claimed in claim 8.

12. A transgenic plant containing a promoter as claimed in claim 1, operatively linked to a gene encoding a protein.

13. A transgenic plant containing a promoter as claimed in claim 2, operatively linked to a gene encoding a protein.

14. A transgenic plant containing a promoter as claimed in claim 3, operatively linked to a gene encoding a protein.

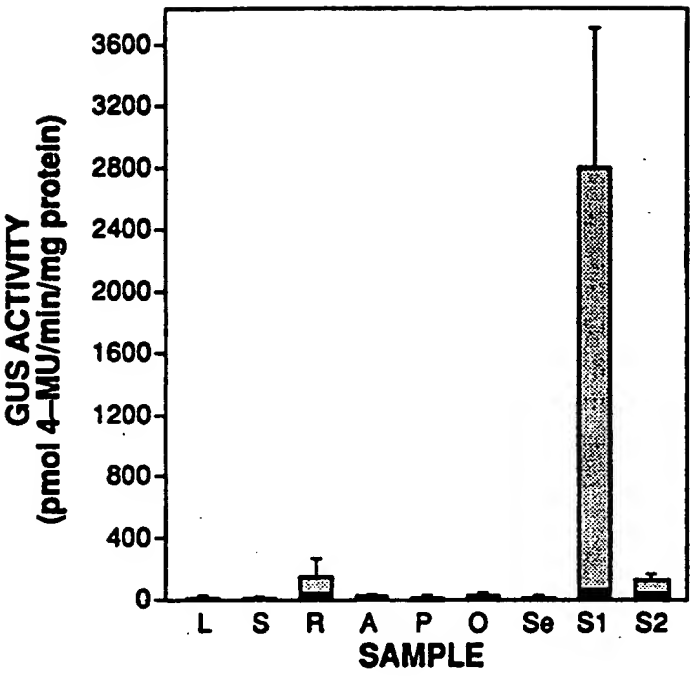


FIGURE 1

Gowling, Strathy & Henderson

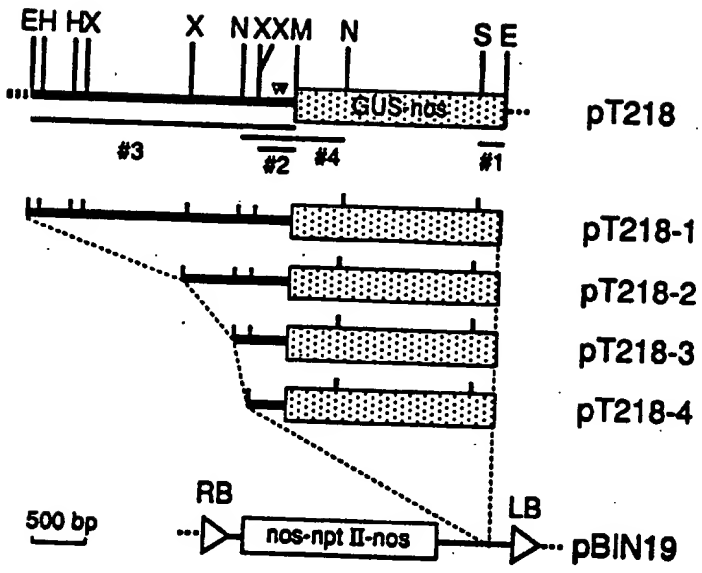

EH HX
X
N XXM
N
S E
GUS-nos
pT218
#3
#2
#4
#1
pT218-1
pT218-2
pT218-3
pT218-4
RB
LB
500 bp
nos-npt II-nos
pBIN19

FIGURE 2

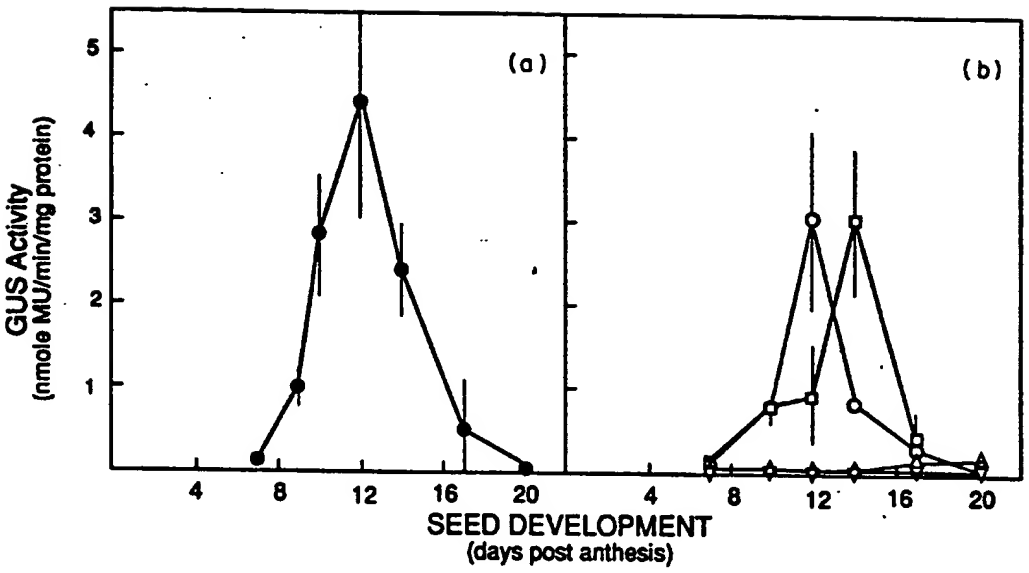

GUS Activity
(nmole MU/min/mg protein)
5
4
3
2
1
(a)
(b)
4
8
12
16
20
4
8
12
16
20
SEED DEVELOPMENT
(days post anthesis)

FIGURE 3

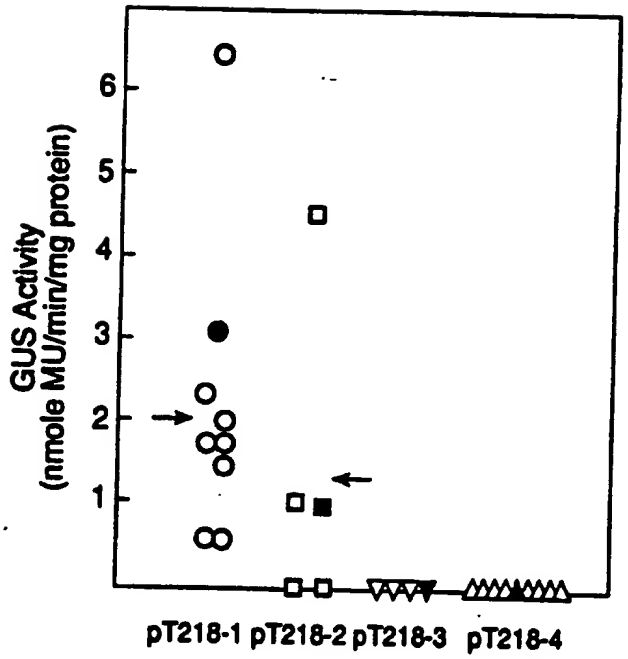

GUS Activity
(nmole MU/min/mg protein)
6
5
4
3
2
1
pT218-1 pT218-2 pT218-3 pT218-4

FIGURE 4

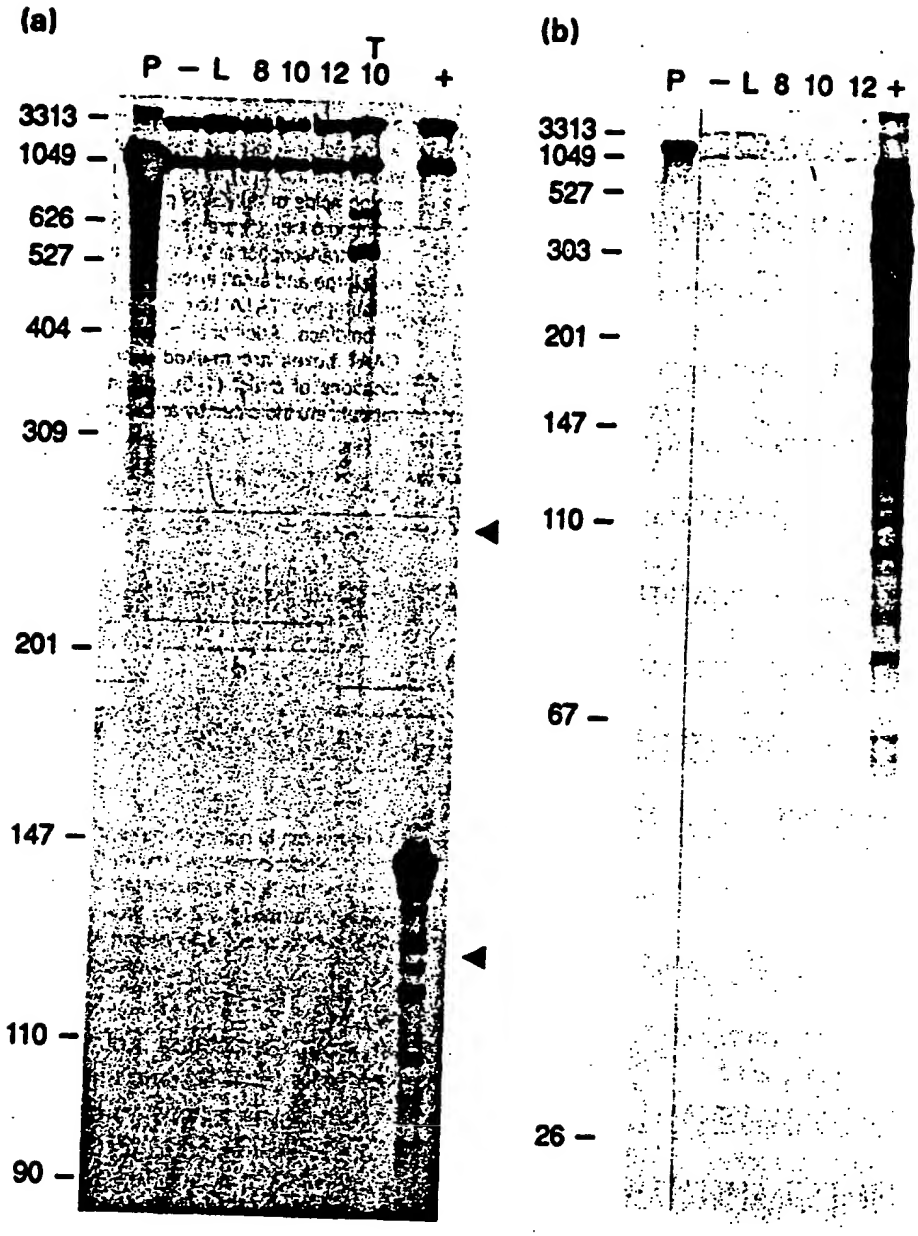

(a)
P
–
L
8
10
12
T
10
+
3313
1049
626
527
404
309
201
147
110
90
(b)
P
–
L
8
10
12
+
3313
1049
527
303
201
147
110
67
26

FIGURE 5

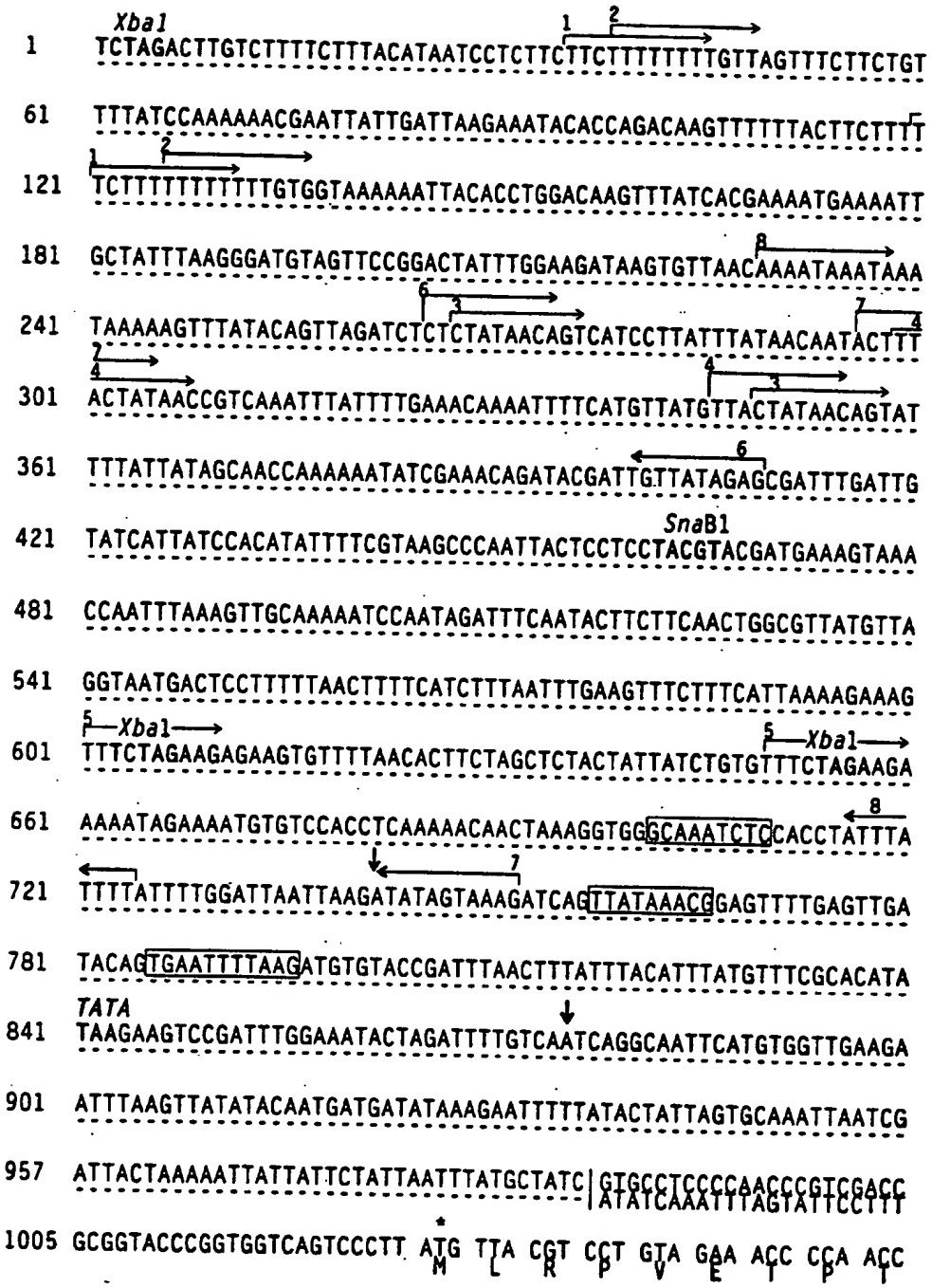

XbaI
SnaB1
TATA
1 TCTAGACTTGTCTTTTCTTTACATAATCCTCTTCTTCTTTTTTTTGTTAGTTTCTTCTGT
61 TTTATCCAAAAAACGAATTATTGATTAAGAAATACACCAGACAAGTTTTTTACTTCTTTT
121 TCTTTTTTTTTTTGTGGTAAAAAATTACACCTGGACAAGTTTATCACGAAAATGAAAATT
181 GCTATTTAAGGGATGTAGTTCCGGACTATTTGGAAGATAAGTGTTAACAAAATAAATAAA
241 TAAAAAGTTTATACAGTTAGATCTCTCTATAACAGTCATCCTTATTTATAACAATACTTT
301 ACTATAACCGTCAAATTTATTTTGAAACAAAATTTTCATGTTATGTTACTATAACAGTAT
361 TTTATTATAGCAACCAAAAAATATCGAAACAGATACGATTGTTATAGAGCGATTTGATTG
421 TATCATTATCCACATATTTTCGTAAGCCCAATTACTCCTCCTACGTACGATGAAAGTAAA
481 CCAATTTAAAGTTGCAAAAATCCAATAGATTTCAATACTTCTTCAACTGGCGTTATGTTA
541 GGTAATGACTCCTTTTTAACTTTTCATCTTTAATTTGAAGTTTCTTTCATTAAAAGAAAG
601 TTTCTAGAAGAGAAGTGTTTTAACACTTCTAGCTCTACTATTATCTGTGTTTCTAGAAGA
661 AAAATAGAAAATGTGTCCACCTCAAAAACAACTAAAGGTGGGCAAATCTCCACCTATTTA
721 TTTTATTTTGGATTAATTAAGATATAGTAAAGATCAGTTATAAACGGAGTTTTGAGTTGA
781 TACAGTGAATTTTAAGATGTGTACCGATTTAACTTTATTTACATTTATGTTTCGCACATA
841 TAAGAAGTCCGATTTGGAAATACTAGATTTTGTCAATCAGGCAATTCATGTGGTTGAAGA
901 ATTTAAGTTATATACAATGATGATATAAAGAATTTTTATACTATTAGTGCAAATTAATCG
957 ATTACTAAAAATTATTATTCTATTAATTTATGCTATC GTGCCTCCCCAACCCGTCGACC
ATATCAAATTTAGTATTCCTTT
1005 GCGGTACCCGGTGGTCAGTCCCTT ATG TTA CGT CCT GTA GAA ACC CCA ACC
M L R P V E T P T

FIGURE 6

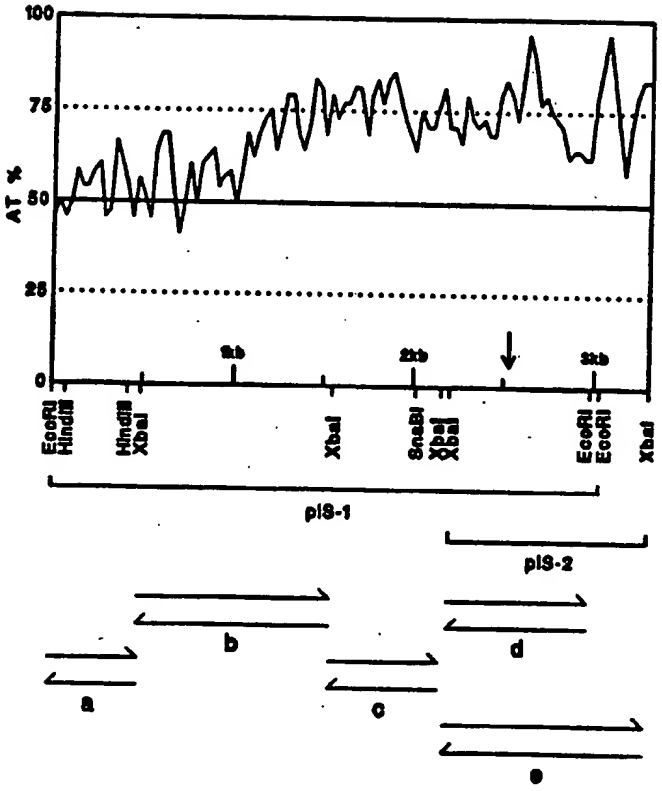


FIGURE 7

Gowling, Strathy & Henderson

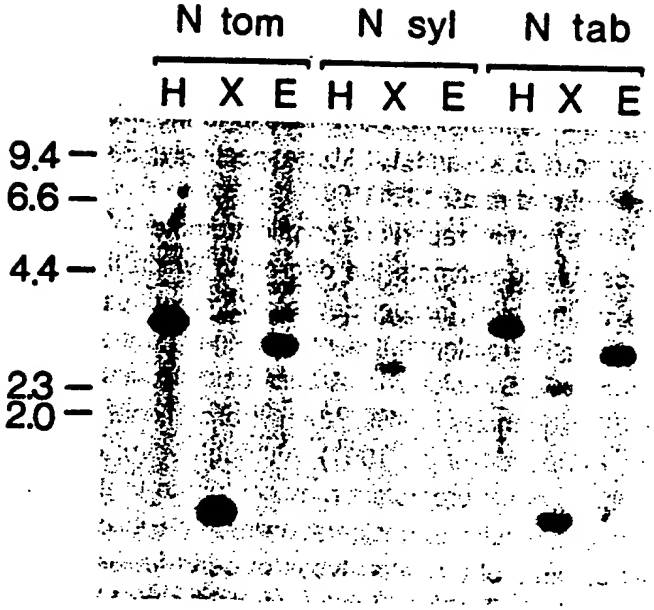


FIGURE 8

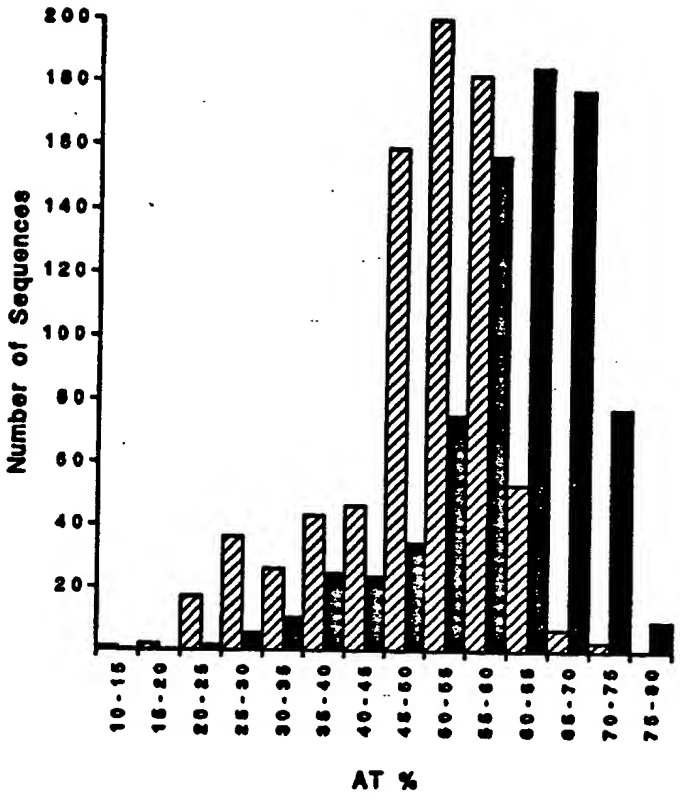

Number of Sequences
200
180
160
140
120
100
80
60
40
20
10-15
15-20
20-25
25-30
30-35
35-40
40-45
45-50
50-55
55-60
60-65
65-70
70-75
75-80
AT %

FIGURE 9